Dipl.-Math. Gordana Djuraš

Generalized Poisson Models for Word Length Frequencies in Texts of Slavic Languages

DISSERTATION

zur Erlangung des akademischen Grades
einer Doktorin der technischen Wissenschaften

TU Graz
Graz University of Technology
Technische Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober
Institut für Statistik

Graz, März 2012

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ....
(Unterschrift)

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other that the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....
date (signature)

Contents

List of Figures

List of Tables

Abstract

This thesis focuses on frequency with which words of a certain length occur in texts, of a particular language, individual author or text type, i.e. *"word length frequencies"*. However, there are many other boundary conditions that have to be taken into account when modeling the word length, such as the choice of an appropriate word definition and the relevant measuring unit for its length. Here, word length is measured in the number of syllables where zero-syllable words, characteristic for the Slavic languages, are considered to be a part of the subsequent word and hence as a separate word class excluded from the analysis.

This study answers the question whether the word length frequency distributions can be theoretically described and if so, is one discrete probability model sufficient to describe them or is more than one model needed. Starting from the Poisson distribution, being a benchmark model for count data, we look for *two-parametric Poisson generalizations* that allow for over-, equi- and underdispersion. Additionally, we require the *1-displaced* and the *size-biased* versions of the obtained distributions to enable modeling when the zero frequency class is omitted. Furthermore, we show that the parameters of theoretical discrete probability models fitted to the observed data play an important role in the attribution of the text types considered.

Three different estimation techniques are compared and the performance of the procedures is evaluated by Monte Carlo simulation studies. Finally, relevant models are applied to real data (120 texts from both Slovenian and Russian languages).

Kurzfassung

Diese Arbeit konzentriert sich auf die Häufigkeit mit der Worte einer bestimmten Länge in Texten einer bestimmten Sprache, eines individuellen Autors oder Texttyps auftreten, d.h. wir diskutieren *Wortlängenhäufigkeiten*. Es gibt aber zusätzliche Randbedingungen, die zu beachten sind, wenn man die Wortlänge modelliert, wie die geeignete Wahl einer Wortlängendefinition und die relevante Masseinheit für die Bestimmung der Länge. Hier messen wir die Wortlänge durch die Anzahl der Silben, wobei null-silbige Wörter, die charakteristisch für slawische Sprachen sind, als Teil des folgenden Wortes aufgefasst und daher als separate Wortklasse von der Analyse ausgeschlossen werden.

Diese Studie beantwortet die Frage, ob empirische Häufigkeitsverteilungen von Wortlängen theoretisch beschrieben werden können, und falls dies möglich ist, ob nur ein diskretes Wahrscheinlichkeitsmodell ausreicht oder ob mehrere Modelle notwendig sind. Ausgehend von der Poissonverteilung als Referenzmodell für Zähldaten, betrachten wir *zweiparametrige Verallgemeinerungen der Poissonverteilung*, die neben dem Referenzfall auch Überdispersion und Unterdispersion abbilden können. Weiters brauchen wir für die Modellierung der Wortlängenverteilung die *1-displaced* und *size-biased* Versionen der Verteilungen, da die Wortlänge zumindest 1 beträgt. Zusätzlich zeigen wir, dass die Parameter der theoretischen diskreten Wahrscheinlichkeitsmodelle, welche an beobachtete Daten angepasst werden, eine wichtige Rolle bei der Zuordnung von Texttypen spielen.

Es werden drei verschiedene Schätzverfahren miteinander verglichen und die Eigenschaften der Methoden durch Monte Carlo Simulationen evaluiert. Schließlich werden relevante Modelle auf reale Daten (jeweils 120 slowenische und russische Texte) angewandt.

Acknowledgement

Despite occasional moments of lack of self-confidence on my part, still many people believed that I would be writing these lines after all. I warmly thank:

- My supervisor Prof. Dr. ERNST STADLOBER for sharing his knowledge with me, guiding me through the work and for dedicating a lot of time, energy and patience to the development of this thesis. Without his tireless and professional support, encouragement and believing in me for years it would be much harder to follow this path to reach the end.

- Prof. Dr. RUDOLF DUTTER, for accepting to be the second referee for my thesis.

- Prof. Dr. PETER GRZYBEK for giving me an opportunity to be a member of the team of FWF project “Word Length (Frequencies) in Slavic Texts” on the Institute for Slavic languages at University of Graz. During the time I spend there, I’ve learned that there is a inseparable connection between linguistics and statistics, although maybe not immediately observable.

- Prof. Dr. HERWIG FRIEDL for suggesting the idea of “size-based distributions”. I am grateful for a helpful advice and interest he showed for my PhD.

- EMMERICH KELIH for his valuable and irreplaceable support in understanding all necessary linguistic terms and definitions; for providing Slovenian and Russian text data and numerous useful linguistic information essential for my PhD; for many hours of exhausting discussions we had on “word lengths”; for reading linguistic parts of my work and for the valuable comments on how to improve them. And after all, for being and staying my friend.

- GABRIEL ALTMANN, for finding the time and interest to discuss both linguistic and statistical topics with me, providing detailed explanations in no time, encouraging me, giving advice and huge support, and showing constant appreciation for my work.

- My friends BIRGIT KORNBERGER and SILVIA LIKAVEC for dedicating many hours to reading my PhD, being always extremely prompt in providing the feedback. Their suggestions helped me a lot to improve the readability of

the work. I am deeply grateful for their help and for all the past years, when they were permanently standing by my side.

∗ GERHARD NEUBAUER, for being a dynamic, passionate and application-oriented co-author. I greatly appreciate numerous productive discussions we had on “discrete distributions”.

∗ ULRIKE KLEB, for encouraging me to find the power I needed to reach the end of this PhD journey and helping me to do the first important steps; for keeping an eye on my progress and for being always ready to discuss with me.

∗ IVONNE SIMON-REITERMAYER for providing me with books and articles from the library of the Institute of Statistics whenever I needed them, and for believing that I will manage to finalize this work. She made the time I spent during my studies on Institute of Statistics unforgettable.

∗ My friend ULRIKE KOHLBACHER who convinced me to come to Graz to pursue PhD and helped me to fight all the cumbersome bureaucracy needed to succeed in it. She made me feel at home here, being someone to rely on.

∗ My parents, mum ŽIVKA and dad DRAGAN, for supporting me and having the strength to let me go away from home, for their unconditional love and the feeling I can always count on them.

∗ My sister DRAGANA for helping me to solve many difficulties along the way, bringing joy and sharing my happiness, being the precious and indispensable part of my life.

∗ My loving husband MIKI for gigantic support, encouragement, and all these years of love, being always an endless source of happiness, comfort, and motivation.

∗ Last but not the least, HERMANN KATZ, the head of Statistical Applications research group, and WOLFGANG POLT, the head of Centre for Economic and Innovation Research at Joanneum Research, for making it possible for me to take ten months of sabbatical to finalize this work.

Foreword

The study of word length can be traced back to more than a hundred years long tradition. Irrespective of the fact that the word, just like the sentence, has a central role for any process of text construction, its length, as a separate theoretical category, has been long neglected in linguistics and literature. Only recent studies dedicated more attention to the question of the frequency with which words of a given length occur in a text, i.e. *"word length frequencies"*. The theory of word length distributions has been developed, providing the systematic evidence that the word length is not a chaotic category, but follows certain regularities. These regularities result from the interaction of diverse *extratextual factors*, such as the concrete language under study, individual authorship, or text type, influencing both word length and word length frequencies. Such external influences should thus be considered when searching for an adequate probability model to describe the observed word length frequencies. The choice of the appropriate model depends, however, also on *theory-driven factors*, in particular the definition of the *word* and the choice of a measuring unit for its length. Yet, there are no uniquely defined categories in language, and thus all linguistic units have to be defined according to the objective of the research.

In Chapter 1, giving briefly the historical development of the word length studies, we discuss some possible operational definitions of the term *word*, favoring, however, different linguistic aspects. The ultimately chosen definition is of relevance for automatic text processing and quantitative text analysis, but also adequate for dealing with questions we are interested in. Moreover, we will show that the choice of a syllable as an appropriate measurement unit for word length leads, especially in Slavic languages, with particular accent on the Slovenian and Russian, to the problem of so-called zero-syllable words. Therefore, it is of utmost importance to make the decision whether to consider or neglect these words as a separate word class, since the resulting probability model depends on this fact. Subsequently, we discuss important steps needed for a systematic and automatic analysis of the word length frequencies for the Slovenian and Russian texts sampled. Lastly, some mathematical notations are introduced.

Starting from the common Poisson model, be the benchmark model for the statistical analysis of the count data, Chapter 2 highlights practical situations where the requirement of the equality of mean and variance is not fulfilled. As a crucial value to detect divergences from the Poisson model the index of dispersion δ (variance to mean ratio) is introduced. According to whether $\delta > 1$, $\delta \approx 1$ or $\delta < 1$, one speaks of Poisson overdispersion, equidispersion and underdispersion, respectively. In order to find a general model for word length frequency distributions, applicable to all three dispersion situations, various approaches for constructing new distributions are considered here. Moreover, two further possibilities for dealing with distributions when the zero frequency class is omitted are given. With a model chosen in this way the next consequence is the estimation of the unknown model parameters. We introduce three common estimation techniques, one of them especially beneficial when the frequency of the lowest class is much higher than the other frequency classes. The algorithm for generating random variables of the discrete distributions considered in the forthcoming chapters is presented, and it is explained how to check whether a certain model may be adequate or not.

Chapter 3 is devoted to Fucks' generalized Poisson distribution and its further generalizations and modifications. Our re-analysis of Fucks' linguistic data provides convincible arguments, that neither the Fucks–Gačečiladze distribution, nor the generalized Poisson model of Fucks, as its special case, can be accepted as an overall theoretical model for word length frequencies, not even for those languages that form their words from syllables, as Fucks himself claimed. Thus, based on this findings we search for other generalizations of the Poisson model.

Since we primarily want to find the simplest model possible, that is both statistically interpretable but also meaningful for linguists, we restrict our further attention to two-parametric Poisson generalizations only. In Chapter 4 the focus is on the main properties of the Singh-Poisson (SP) distribution, whereas Chapter 5 describes in detail the Hyper-Poisson model, frequently used for word length distributions but also as a model of the distribution of the sentence length. Difficulties including the computation and derivation of the confluent hypergeometric function are discussed and problem-solving approaches are presented. The performance of the estimation procedures is evaluated for both models by Monte Carlo simulation studies.

The generalized Poisson (GP) model, discussed in Chapter 6, approximates both the negative binomial and binomial model, being thus suitable for over- and underdispersed count data. However, it allows also for modeling equidispersed data. One of its two parameters measures dispersion, hence tunes the type of the distribution. Based on simulation studies we evaluate whether the GP model or the SP model perform better under the three dispersion situations. Because both distributions are two-parametric we fix the first two moments to obtain meaningful comparison.

Chapter 7 provides evidence, that the Cohen-Poisson distribution, belonging to the class of misrecorded Poisson distributions, is not a reasonable theoretical model for the texts studied. The arguments are based on the index of dispersion.

The thesis ends with Chapter 8 applying the relevant models to 120 Slovenian as well as Russian texts selected and presenting the main conclusions of the study for both Slavic languages.

Graz, March 2012 Gordana Djuraš

Chapter 1

Introduction

1.1 Statistics Meets Linguistics

This study is an attempt to bridge the gap between humanities and natural sciences. The starting point is linguistics, a science dealing with all aspects of describing and explaining the nature of human language. It basically focuses on the analysis of language structure, grammar, formation and/or modification of words and phrases, and composition of grammatical sentences from these words. Other fields of linguistics explore e.g. the meaning of words, language change over time, its history and the use of language in texts. Quantitative linguistics in particular endeavors to detect, describe and explain the underlying linguistics rules.

There are more than 5500 spoken languages in the world. Many languages use words as their basic elements. Words are joined together in sentences that can be long or short, simple or complex. Sentences put together create a text. Hence, at the first sight, a text looks like a picture drawn by words. It may seem that these words appear in a text chaotically, without any systematical order, but empirical results show that these words follow a certain regularity.

We cross the borders of linguistics and turn our attention to statistics, looking for the laws responsible for text generation. The question of frequency with which words of a certain length occur in texts of a given language, of a given author or of a given genre has raised interest among linguists. A new theory of word length distributions has been developed. The word length and its frequency, as explanatory variables, and the frequency distribution are in the center of our attention. This study provides theoretical insight into the structural rules of text construction and processing, trying to answer what is actually happening in the language and text.

Very early studies on word length frequency distribution assumed that there might be a unique model for at least all those languages, which form their words from syllables. This work has indeed an intention to show that a single model is not sufficient to cover the wide spectrum of different texts from different languages. Even more, texts from one single language but of different genres often show significant differences in their word length distributions. Therefore, we are trying to distinguish influences linked to the specific language and the ones beyond it. This thesis gives answers to the following questions:

- Can the word length frequency distributions of our sampled texts be theoretically described, and if so, is one discrete probability model sufficient to describe them, or is more than one model needed?
- If more than one model is necessary for the description of a particular text sample, it may be beneficial to establish the connections between these models. Even more, our interest is to find out whether models that come into question can be derived as special cases (or sub-models) of one complex, unifying model. Consequently, the differences that appear between models could be explained through specific changes in parameter values.
- What is the influence of individual factors, such as text type or author on word lengths and word length frequency distribution?
- Based on the answers to the questions above, it is further important to find out whether one can discriminate texts by using the parameters of the given model(s) as discriminant variables. In case of a positive answer, this would give us the possibility to assign a certain text to a text group by classifying the parameter values of the fitted model.

Looking for a suitable theoretical model to explain observed word length frequencies, our aim is primarily to find the simplest model possible, i.e., a model with a minimal number of parameters which is both statistically interpretable but also meaningful for linguists and thus can be used for the characterization of texts and language(s). In trying to find such a model we start from the Poisson distribution being the simplest and most widely used one when modeling count data and searching for the solution within the Poisson family. The statistical analysis is based on texts from two particular Slavic languages, namely Slovenian and Russian.

1.2 Historical Background

Throughout history the problem of modeling the distribution of word length was not only the interest of linguists, but also of scientists from other areas such as physics, mathematics and statistics. In 1851 the English mathematician and logician Augustus De Morgan was the first to point out the relevance of the length of a linguistic unit. He mentioned word length as a possible style characteristic and a useful indicator for identification and discrimination of authorship. In a letter to a friend, he suggested to use the criterion of word length to find out who the real author of the St. Paul's Epistles is (cf. Lord, 1958). Moreover, concentrating on the word length measured by the number of letters, De Morgan postulated that the mean word lengths in two texts written by the same author, even though on two different subjects, are much closer to each other than in two texts written by two different persons on the same topic.

Several other scientists have dealt with the same topic counting even the frequency with which words of a given length occur in a text. The American physicist and meteorologist T. C. Mendenhall (1887) tried to identify the author of a certain composition through the frequency distribution of words of different length. Using preferably graphical methods to represent results obtained, he assumed that "characteristic curve" or "word spectrum" of any composition of the same author does not differ essentially, indeed it distinguishes itself from the curve of another writer. For that reason, he believed, his "word spectrum" method could be applied in cases of disputed authorship. Comparing the word length frequency distribution of the works of Shakespeare and Bacon, Mendenhall (1901, p. 104) found out that his result did not support long-standing claims that Bacon had been the true author of the works usually attributed to Shakespeare. Being aware of the fact that variation can occur "between authors" but also "within authors" Williams (1940, 1956) indicated the necessity to study many different works of the same author, written in different styles, on different subjects and at different periods of his life, to find out to what extent the author defines his characteristic distribution. Furthermore, Williams (1975) showed that Mendenhall failed to consider "genre differences" that could invalidate his conclusions. His numerous empirical investigations made one incontrovertible fact very clear: the word length is not only influenced by the individual style of an author, but may also dependent on diverse other factors, genre being one of them (cf. Grzybek et al., 2005b; Kelih et al., 2005).

The first probability model concerning word length studies was constructed in the 1940s. Observing the distribution of word length measured by the number of syllables, S. G. Čebanov (1947), a Russian military doctor, found the Poisson distribution to be the most appropriate general model for texts of the Indo-European group of languages (cf. also Best and Čebanov, 2001, p. 282; Best, 2005, p. 261; Grzybek, 2006, p. 26). Independently, the German physicist Wilhelm Fucks (1955, 1956a, 1956b) came to a similar conclusion. He aimed at a mathematical description of word formation through syllables by introducing a mixture of Poisson probabilities, known as Fucks' Generalized Poisson distribution (cf. Fucks, 1956a, 1956c). The Poisson distribution is a special case of this general distribution model. Under particular conditions also the Dacey-Poisson distribution can be derived as a two-parameter special case of the proposal of Fucks (cf. Antić et al., 2005)[1]. Fucks' work was extremely influential on the studies done from the 1950s until the late 1970s in the field of quantitative linguistic, in particular on word length. His theoretical assumptions became a starting point for further generalizations, alternative parameter estimation approaches and extensions to other languages. Just a few years later, Cercvadze, Čikoidze, and Gačečiladze (1959) applied Fucks' ideas to Georgian linguistic material by generalizing once more his proposed model (cf. also Gačečiladze

[1] Antić is the maiden name of Djuraš.

and Cilosani, 1971). Two Polish authors, Bartkowiakowa and Gleichgewicht (1964, 1965), suggested an alternative way to estimate the parameters of the Fucks' distribution. Also worth mentioning is the contribution of Vranić and Matković (1965). They suggested the modification of the Poisson distribution as a solution for existing discrepancy in monosyllabic and disyllabic words (cf. also Vranić, 1965).

An intensive examination of word length began with Grotjahn's (1982) modification of Fucks' approach. There are several factors existing in the mind of the author that control the flow of information in a text and the text generation at all. They have influence not only on individual words but also on entire length classes. With this in mind, Grotjahn argued that the probability of word appearance might change in the text due to some factors such as context, with a change of topic or after an interruption of writing. He proposed mixtures of distributions where the parameter of the Poisson distribution is considered as a random variable following the gamma distribution. The resulting marginal model is the well-known negative binomial distribution and provides a good fit, as Grotjahn showed, at least for German texts. Grotjahn's contribution was an important step in the history of word length studies. He showed that instead of searching for one general model for all languages and all texts, one should rather concentrate on a variety of distributions.

Afterwards Wimmer et al. (1994) introduced an entirely new concept of word length frequency distributions. The distribution of word length is a result of complex processes and mechanisms which need not lead to one single pattern. The author produces a text without caring about the length of a chosen word, since the primary focus is on the text content. Moreover, if e.g. the author takes a brake during the writing and starts in a different mood the conditions may change. As a consequence, e.g. in long chapters of a novel, word length may show irregular patterns due to these structural breaks. Hence, as already stated by Grotjahn and Altmann (1993), quite rarely one single model can be sufficient for the same language, even by the author himself, some modifications are desirable. Searching for regularities which might affect word length distribution Wimmer et al. (1994, p.101) assumed that the various word length classes do not evolve independently of each other. Their flexible system of distributions is based on the idea that each two neighboring probability classes are proportional, i.e. $P_x \propto P_{x-1}$. The proportionality results from the interaction of a large number of different factors, such as author, genre, language, epoch, etc., responsible for the text generation. These should be considered when searching for the adequate model solution. In order to construct a model as universal as possible in its applicability, i.e. valid for all texts, all authors, all genres and all languages, the following assumption is made. The proportion of word length classes is not constant, rather it can be understood as a function of length, denoted by $g(x)$. Consequently, the word length distributions are generated by the basic mechanism defined by the formula

$$P_x = g(x)P_{x-1}\,, \tag{1.1}$$

where P_x denotes the probability of the word length x and P_{x-1} is the probability of the word length $x - 1$. This approach has the advantage that in order to determine some particular distribution, it is not necessary to know the probabilities of all individual frequency classes. A special choice of the function $g(x)$ leads to different models. The simplest function $g(x)$ has the form of Menzerath's law[2], i.e.

$$g(x) = ax^{-b}\,, \;\; a > 0,\, b \geq 0\,. \tag{1.2}$$

Wimmer et al. (1994, 101ff.) distinguish three levels of modeling the distribution of word length: (i) elementary form, (ii) modifications, and (iii) generalizations.

(i) The most elementary form of a word length distribution arises by setting the function $g(x)$ given in (1.2) into the formula (1.1). Hence, we have

$$P_x = \frac{a^x}{(x!)^b} P_0\,, \;\; x = 0, 1, 2, \ldots, \;\; a,\, b > 0\,, \tag{1.3}$$

where P_0 is the normalization constant which can be computed from $\sum_x P_x = 1$. This finally results in the *Conway-Maxwell-Poisson distributions* (cf. Wimmer et al., 1994; Wimmer and Altmann, 1996)

$$P_x = \frac{a^x}{(x!)^b T_0}\,, \;\; x = 0, 1, 2, \ldots, \;\; a, b > 0, \; T_0 = \sum_{j=0}^{\infty} \frac{a^j}{(j!)^b}\,. \tag{1.4}$$

These distributions appear typically in Slovak poetry texts (cf. Nemcová and Altmann, 1994, 215f.).

(ii) On the second level of modeling we distinguish between local and global modifications, also called "first order extensions" (cf. Wimmer and Altmann, 1996). Local modifications influence only probability classes and they can be of two kinds: "unrestricted" and "restricted". "Unrestricted" modifications occur as a result of displacement in the given language, such as the increase of zero-syllable[3] words in Slavic languages. Here all frequency classes are modified by scalars whose sum is 1. "Restricted" modifications affect only some frequency classes and are mostly influenced by personal fluctuations in texts (cf. Wimmer and Altmann, 1996, p. 113).

A global modification occurs when the parameter b in (1.2) is set to 0 or to 1. Setting $b = 0$ yields $g(x) = a$ which for $0 < a < 1$ results in the *geometric distribution*. In this case the proportionality is constant. Its linear extension $g(x) = a(R - x + 1)$ gives the *Palm-Poisson distribution*

$$P_x = \frac{R_{(x)} a^x}{\sum\limits_{j=0}^{R} R_{(j)} a^j}\,, \;\; x = 0, 1, \ldots, R\,, \tag{1.5}$$

[2] The general formulation of Menzerath's Law says: "The longer a language construct the shorter its components (constituents)". According to this hypothesis, e.g. the length of syllables, $g(x)$ is inversely proportional to the length of the words, x in which they occur. It means, "the longer the word the shorter its syllables" (cf. Altmann, 1980, p. 1).

[3] For the explanation of zero-syllable words see Section 1.5.

with $R_{(x)} = R(R-1)\ldots(R-x+1)$, $R \in \mathbb{N}$, $a > 0$, which proved to be useful for Italian texts (cf. Altmann et al., 1997).
The choice $b = 1$ is more frequent and implies $g(x) = a/x$, wherefrom we easily obtain the *Poisson distribution* with parameter a. In German language where zero-syllable words do not exist, the positive (zero-truncated) Poisson distribution can be a useful model (cf. Altmann and Best, 1996). Its unrestricted local modifications result in the *positive Pandey-Poisson*, *positive Singh-Poisson*, and *positive Cohen-Poisson distributions* (cf. Wimmer and Altmann, 1996, 117ff.). Furthermore, four linear extensions of the function $g(x) = a/x$ are possible:

(a) from $g(x) = (a+bx)/x$ with reparametrisation $a/b = k-1$, $b = q$ we obtain the negative binomial and when there are no zero-syllable words the positive *negative binomial distribution* which proved to be adequate for German texts (cf. Altmann and Best, 1996),

(b) the choice $g(x) = (a-bx)/x$ after reparametrisation $a/b = n+1$, $n \in \mathbb{N}$, $b/(b+1) = p$ yields the binomial and when there are no zero-syllable words the *positive binomial distribution*, typical for Turkish texts (cf. Altmann, Erat, and Hřebíček, 1996). In Polish texts, where the zero-syllable prepositions are treated as a separate closed class of zero-syllabic words, the positive binomial distribution upgraded for the class $x = 0$ results in the *extended positive binomial distribution* (cf. Uhlířová, 1996, 1997),

(c) taking $g(x) = a/(b+x-1)$ leads to the *Hyper-Poisson distribution*, which can be found e.g. in Italian (cf. Altmann et al., 1997) and Slovak journalistic texts (cf. Nemcová and Altmann, 1994),

(d) for $g(x) = (a+bx)/(c+x)$ after substitution $a/b = k-1$, $b = q$, $c = m-1$ we obtain the *Hyper-Pascal distribution* which turned out to be appropriate, e.g. for Slovak prose texts (cf. Nemcová and Altmann, 1994).

For illustration purposes we present the derivation of the negative binomial distribution. Using reparametrisation given in (a) and after some calculations we have

$$P_x = b\frac{\frac{a}{b}+x}{x}P_{x-1} = q\frac{k+x-1}{x}P_{x-1} = q^x\frac{(k+x-1)!}{x!(k-1)!}P_0 = q^x\binom{k+x-1}{x}P_0\,.$$

Since,

$$1 = \sum_{x=0}^{\infty} P_x = P_0\sum_{x=0}^{\infty} q^x\binom{k+x-1}{x} = P_0\sum_{x=0}^{\infty}(-q)^x\binom{-k}{x} = \frac{P_0}{(1-q)^k}$$

we obtain $P_0 = (1-q)^k = p^k$.

(iii) On the third level further generalizations are possible (cf. Wimmer et al., 1994; Wimmer and Altmann, 1996; Altmann, 2005). These more complex models assume that the set of word length classes is organized as a whole. It means that the class x is controlled not only by the neighboring class $x-1$, but also by all j previous classes, yielding

$$P_x = g(x) \sum_{j=1}^{x} h(j) P_{x-j}\,, \quad j = 1, 2, \ldots, x\,, \tag{1.6}$$

where $h(j)$ denotes the weighting function that can itself be a probability distribution and $g(x)$ is the proportionality function as in (1.1). The probability of class x results as a weighted sum of all lower classes. Obviously, (1.1) is a special case of (1.6). Wimmer et al. (1994, p. 103) considered (1.6) to be a general proportionality form wherefrom various word length frequency models for all languages can be derived. The correct specification of the functions $g(x)$ and $h(j)$ is required. If we choose e.g. $g(x) = a/x$ and $h(j) = jH_j$, where H_j is itself a probability function of a random variable J, then we get the class of well-known generalized Poisson distributions. Setting H_j to have Borel distribution (cf. Johnson et al., 1992, 394f.)

$$H_j = \frac{j^{j-2} b^{j-1} e^{-bj}}{(j-1)!}\,, \quad j = 1, 2, \ldots\,, \quad 0 \le b < 1 \tag{1.7}$$

results in the *Consul-Jain-Poisson distribution*, also known as Consul's generalized Poisson distribution[4] (cf. Wimmer and Altmann, 1999, p. 93)

$$P_x = \frac{a(a+bx)^{x-1} e^{-(a+bx)}}{x!}\,, \quad x = 0, 1, \ldots\,, \quad a > 0\,,\ 0 \le b < 1\,. \tag{1.8}$$

Lately, Wimmer and Altmann (2005) presented an approach that unifies all previous assumptions, linguistic hypotheses and empirical findings. In order to construct an overall discrete model they looked for requirements that can affect the behavior of the relative rate of change of the probability of the word length x, i.e.

$$\frac{\Delta P_{x-1}}{P_{x-1}} = \frac{P_x - P_{x-1}}{P_{x-1}}$$

and obtained the general recurrence formula

$$P_x = \left(1 + a_0 + \sum_{i=1}^{k_1} \frac{a_{1i}}{(x-b_{1i})^{c_1}} + \sum_{i=1}^{k_2} \frac{a_{2i}}{(x-b_{2i})^{c_2}} + \ldots\right) P_{x-1}\,, \quad c_i \ne c_j\,,\ i \ne j\,. \tag{1.9}$$

Various well-known families of distributions can be derived from this general formula. After the substitution $k_1 = k_2 = \ldots = 1$, $a_{11} = a_1$, $b_{11} = -b_1$ and $a_{21} = a_{31} = \ldots = 0$ we obtain

$$P_x = \left(1 + a_0 + \frac{a_1}{(x+b_1)^{c_1}}\right) P_{x-1} \tag{1.10}$$

representing the family of word length distributions frequently used for Slavic languages (cf. Rottmann, 2006). Obviously, this simplified recurrence formula is a

Table 1.1: Some distributions used in word length studies of Slavic languages

Name	Conditions on parameters	$g(x)$	P_x
Conway-Maxwell-Poisson	$a_0 = -1$, $a_1 = a$, $b_1 = 0$, $c_1 = b$	ax^{-b}	$\frac{a^x}{(x!)^b} P_0$
Poisson	$a_0 = -1$, $a_1 = a$, $b_1 = 0$, $c_1 = 1$	$\frac{a}{x}$	$\frac{e^{-a}a^x}{x!}$
binomial	$a_0 = -1 - b$, $a_1 = a$, $b_1 = 0$, $c_1 = 1$	$\frac{a - bx}{x}$	$\binom{n}{x} p^x q^{n-x}$
negative-binomial	$a_0 = -1 + b$, $a_1 = a$, $b_1 = 0$, $c_1 = 1$	$\frac{a + bx}{x}$	$\binom{k + x - 1}{x} p^k q^x$
Hyper-Poisson	$a_0 = -1$, $a_1 = a$, $b_1 = b - 1$, $c_1 = 1$	$\frac{a}{b + x - 1}$	$\frac{a^x}{b^{(x)} {}_1F_1[1; b; a]}$
Hyper-Pascal	$a_0 = -1 + b$, $a_1 = a - bc$, $b_1 = c$, $c_1 = 1$	$\frac{a + bx}{c + x}$	$\frac{q^x k^{(x)}}{m^{(x)} {}_2F_1[k, 1; m; q]}$

special case of the basic mechanism in (1.1) where $g(x) = 1 + a_0 + a_1/(x + b_1)^{c_1}$. Table 1.1 gives an overview of some frequently used word length distributions for Slavic languages. These can be derived from the recurrence formula (1.10) under appropriate conditions listed in Table 1.1 [5] and above given reparametrisations.

The developed theory shows that there is a general mechanism behind various language processes represented by (1.9). Still, there is no explanation why a particular language prefers one special model and which language phenomena affect the values of the parameters of appropriate models.

1.3 The Problem of Representativeness

Let us assume that there is a universe of texts consisting of an infinite/finite number of textual objects. To explain the structure of this universe it is first important to define a *text* as an analytical unit and then to classify these objects. The process of classification includes identification and description of hierarchically ordered subsystems and has to be realized with respect to theoretical assumptions. Some textual objects are more similar to each other, some of them differ instead. Therefore, any

[4] Detailed analysis of this distribution is given in Chapter 6.

[5] In Table 1.1 $b^{(x)}$ denotes the Pochhammer's symbol, sometimes called the rising (or ascending) factorial, ${}_1F_1$ is the confluent hypergeometric function, and ${}_2F_1$ is the Gaussian hypergeometric function.

kind of classification involves also quantification. To apply statistical methods a basic theoretical prerequisite to be fulfilled is the homogeneity of the texts examined. Closely associated with homogeneity is the question of a *representative sample* that should reflect the characteristics of the whole textual *population.* However, the conventional interpretation of the population of a particular language as a collection of a variety of different texts written by different authors and text types, is not plausible. In fact, such an approach, though striving to achieve "representative" language description, leads to a mixture of heterogeneous text material, of a "quasi" text, in a way, for which harmonious behavior of observed frequencies cannot be expected (cf. Kelih, Buk, Grzybek, and Rovenchak, 2009). Notice that any corpus is one such mixture. On the contrary, quantitative text analysis focuses explicitly on individual texts endeavoring to minimize "internal" heterogeneity, characteristic for corpus analysis (cf. Altmann, 1992). As pointed out by Kelih (2009a) particular regularities hold primarily for semantically coherent, holistic and closed linguistic texts. These quantitative regularities, not necessarily present in text segments, are likely to interfuse in any kind of text mixture (cf. Grzybek et al., 2005b). Thus, following the approach given in Strauss, Grzybek, and Altmann (2006), chapters of a longer prose text (such as a novel) will be treated as separate analytical units, whereas complete poetic, journalistic texts and letters are considered as homogeneous individual texts.

In the quantitative approach, as mentioned in Grzybek et al. (2005b), text universe is structured with respect to the concepts of *functional styles* and *text types* (cf. Figure 1.1). Commonly, five to eight different functional styles are considered, such

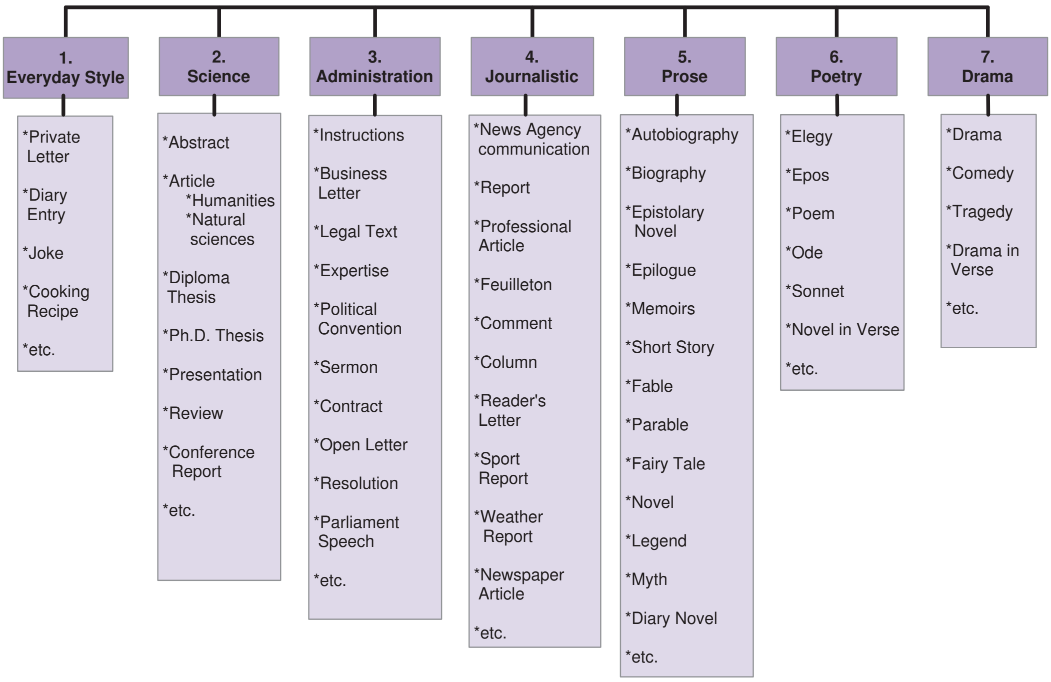


Figure 1.1: Functional styles and text types

as everyday, official/administrative, scientific, journalistic or artistic style. However, such categorization results in an extreme heterogeneity of the texts attributed to the individual categories. The concept of text sorts describes the whole spectrum of text types as a population of ca. 4000 various text types. Clearly, it is impossible and even not necessary to take into consideration all of them. For the systematic statistical analysis it should be rather guaranteed that the whole spectrum of different stylistic shapes is correctly represented (cf. Kelih et al., 2009). Consequently, based on above considerations and intensive empirical work on word length frequencies in Slavic languages (cf. Kelih et al., 2005; Grzybek et al., 2005b; Grzybek, Kelih, and Stadlober, 2005a), it proved to be efficient to combine these two principles. In order to cover the broad textual spectrum, further quantitative and statistical analysis should be based on the scheme presented in Figure 1.1. The selection of individual texts will be with regard to their attribution to a particular text type and functional style. In this respect, an individual text represents a larger group of similar texts. For example, a particular poem is a representative of the "population" of all poems, hence based on the knowledge derived from its properties (e.g. word length) we can make conclusions about the characteristics of the entire population.

Finally, in addition to the decision made above, it still remains to be defined what a "word" is and in which units word length should be measured. This topics will be considered in the following sections.

1.4 On the Definition of Word and Word Length

The topic of our interest is a text, a written sequence of statements which consists of graphemes, syllables, words, sentences or even elements of higher order such as paragraphs, sections and chapters. Thereby, a *word* has a central role in linguistics. It seems that in everyday life, everyone has an idea of what the term word implies. Yet, there is no generally accepted definition of this term, not even in linguistics[6]. The basic requirement for the quantitative study of the word length is to define all linguistic units that are subject of quantification, thus also a *word* has to be defined according to the objective of the research. Knowing that there is no uniquely accepted, general definition which we can use for our purposes, we should first discuss some available definitions and then choose one, adequate for dealing with concrete questions we are interested in.

Within the framework of quantitative linguistics, the mostly used operational definitions of a *word* are the following three alternatives, discussed in detail in Kelih (2007) and explained here on the example of a Russian sentence:

(a) *orthographical* - describes a word as a unit of a text which appears as a sequence of written letters between spaces (cf. Bühler et al., 1972; Bünting

[6] There are over 200 different definitions of the term *word* in linguistics. A detailed overview is given in Krámský (1969). For example in a written German text every sequence of letters between two blanks can be considered as one word.

and Bergenholtz, 1995). Such a definition, applicable only to written texts, has been fundamentally criticized by many linguists. Some of the arguments against this definition are: (i) some languages have never been alphabetized, (ii) the space and other punctuation marks do not have a word-separating function in all languages, as e.g. in Chinese where text appears as a chain of characters strung together and (iii) the space must not be generally considered as a reliable and consistent means of separating words. Nevertheless, for languages with a long "writing" tradition this definition can be used as a starting point.

|Он| |вслущивался| |в| |трубку|, |волнуясь| |и|
He *listen* *into* *headset* *excited* *and*

|стараясь| | уловить| |что| | то| .
trying *to understand* *something* –

(b) *orthographic-phonetic* - defines a word, based on an orthographic criterion, as a unit between spaces where zero-syllable words (for explanation see Section 1.5), both not stressed and proclitically[7] or enclitically attached to the following or previous word respectively, are to be treated as a part of a neighboring word (cf. Antić et al., 2006a). According to this definition zero-syllable words in the word length studies on Slavic languages do not exist as a separate word class, rather they "merge" with other full-accented words in the process of word length determination.

|Он| |вслущивался| |**втрубку**| [8] ,|волнуясь| |и| |стараясь| | уловить|
|что| | то|.

(c) *phonological* (from *phono = concerning sound*) - takes an independent, accentuated unit as a word, a unit that sounds like one word when spoken. It means, the word is defined as so-called "rhythm group". In contrast to the orthographic-phonetic definition, not only zero-syllable pro- and enclitic words, but also all potentially not accentuated words are taken into consideration. As mentioned in Kelih (2007), in Russian these groups of words without an accent include primarily prepositions (на, из, за, о, без, дљa), conjunctions (such as и, а, но, ни), as well as interjection and particles (вот, же, ни, не, ли).

|Он| |вслущивался| |**втрубку**|, |волнуясь| |**истараясь**| [9] |уловить|
|**чтото**|[10].

[7] A clitic is a morpheme that has syntactic characteristics of a word, but shows the evidence of being phonologically bound to another word. A clitic that is attached to the beginning (end) of another word is called proclitic (enclitic).

[8] в has no accent, thus according to Lehfeldt (1999) is a proclitic and merges with трубку

[9] и has no accent, hence forms proclitically one word with стараясь

[10] то is a particle with enclitic behavior to word что

Without a doubt, each of the three described definitions favors different linguistic aspects and causes significant differentiation in statistical measurements as clearly illustrated in Table 1.2. The number of words and the first two empirical moments of the above displayed Russian sentence vary with respect to diverse word definitions, while the number of syllables stays unmodified. Consequently, this results in the necessity to introduce different statistical models. The ultimately chosen definition should therefore be of relevance for automatic text processing and quantitative text analysis.

Table 1.2: Statistical measurements due to two different *word* definitions

Definition	Number of words	Number of syllables	Mean word length	Standard deviation
orthographical	10	19	1.90	1.28
orthographic-phonetic	9	19	2.11	1.16
phonological	7	19	2.71	1.11

Apart from the linguistic difficulty to define the concept of a *word*, data construction in the length research is associated with the problem of the *measurement unit*. The word length is usually measured in the number of components of lower-level units which in turn have their own lengths. Grotjahn and Altmann (1993, p. 142) stated that there are three basic types of units to measure the word length, namely (a) graphical, (b) phonetic, and (c) semantic ones. In most of the languages the graphical units are letters. A *letter* is the smallest unit that can be used to measure the word length. Sounds, phonemes or syllables are phonetic units, while morphemes are the semantic measurement units. A *morpheme* is the smallest linguistic unit that has a meaning. In a spoken language, a morpheme is composed of *phonemes*, the smallest linguistically distinctive units of sound and in written language it consists of *graphemes*, the smallest units of written language. As can be seen, even the measurement unit is not uniform and naturally given, it must be established for each aim to be achieved and the resulting mathematical model (inclusively the theory) depends always on the choice of the measurement unit. This fact is clearly demonstrated by Figure 1.2 which presents the distribution of word length measured in letters and syllables in Ivan Cankar's letter to Ana Lušinova. Obviously, if word length is measured in letters we obtain much more frequency classes, and the heterogeneity structure requiring quite complex theoretical models. In contrast, when modelling the distribution of word length measured in syllables, simple and easy interpretable models are sufficient. Moreover, in quantitative analysis of word length in texts, a word is usually measured by the number of syllables, since the syllable is considered to be a direct constituent of the word (cf. Altmann et al., 1997, p. 2). By definition, a syllable is any of the units into which a word can be divided, usually consisting of a vowel sound (or at least one "syllabic" segment) with a consonant before and/or after it. Therefore, in order to automatically measure word length it is not primarily necessary to define syllable boundaries; rather it is sufficient to count the number of vowels per word. Other measurement units such as phonemes or morphemes are also possible, but make the analysis of quantitative properties of a language pretty complicated.

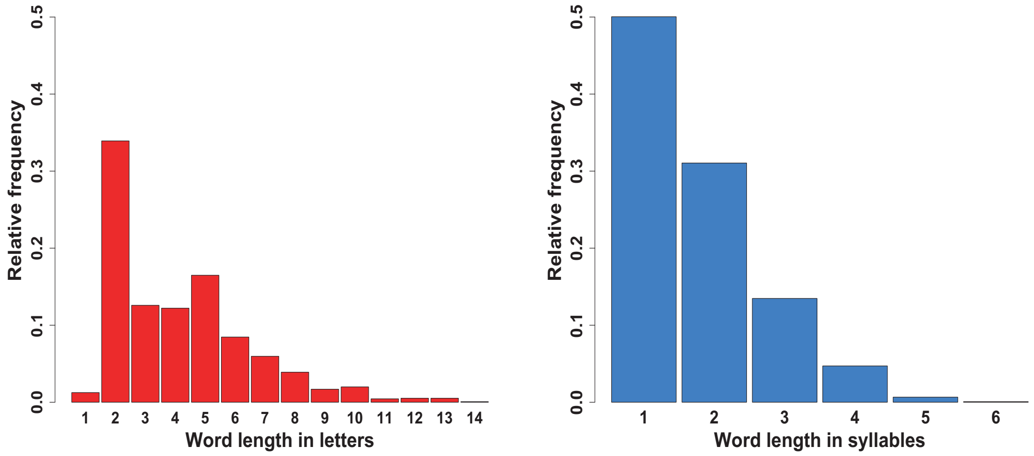


Figure 1.2: Distribution of word length measured in letters (left) and syllables (right)

In the present work, due to the analytical unit, the word length is to be measured in the number of syllables per word, since the number of syllables can be easily determined and automatically calculated, at least for Slovenian and Russian texts which represent the text material of the present study.

1.5 Zero-Syllable Words in Slavic Languages

The above mentioned definition of the word as an orthographically defined unit (every chain of letters separated by blanks is considered to be a word) and a choice of syllable as an appropriate measurement unit (as direct constituent of the word) lead, especially in Slavic languages, to the problem of so-called zero-syllable words. In general, these words are prepositions which neither exhibit syllabic vowels nor an independent accent and may thus be treated as clitics from a phonetic viewpoint. Therefore, a further important step for the quantitative study on word length is to decide whether these zero-syllable words should be treated as a separate word class or not. Figure 1.3 shows the distribution of word length in a letter of I. Cankar to A. Lušinova (see Appendix A, Table A.1, text no. 18) for both cases. Obviously, the decision to consider or neglect this specific word class will result in different probability models.

The problem of zero-syllable words has become quite evident in a number of word length studies in Slavic languages when modelling the word length frequency

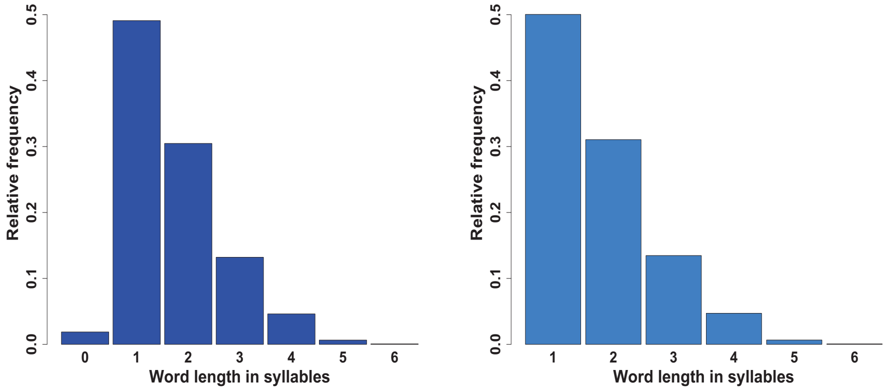


Figure 1.3: Distribution of word length measured in syllables (left: with zero-syllable words, $n = 1385$, $\bar{x} = 1.718$, $s = 0.935$; right: without zero-syllable words, $n = 1359$, $\bar{x} = 1.751$, $s = 0.913$)

distributions. With regard to this question two different approaches exist. On the one hand, there are studies where the zero-syllable words have been treated and analyzed as a separate word class, such as in works of Nemcová and Altmann (1994) in Slovak, Uhlířová (1996, 1997) in Czech or Girzig (1997) in Russian. On the other hand, some researchers favored the orthographic-phonetic definition which negates zero-syllabic words as a category on its own. For example, Wilson (2006), who analyzed Lower Sorbian[11] newspaper texts, argued that in case these prepositions are counted as independent words only rare probabilistic models can be fitted to the data. Since these zero-syllable words do not contain any vowels Wilson claimed that, in accordance with the principles of the Göttingen project[12], they should be treated as a part of the neighboring words (i.e. as clitics) and disregarded from the analysis. As a result more regular probability models can be obtained (cf. Figure 1.3). In analogy to the other Slavic languages, a special problem of a class of zero-syllable words appears also in Slovenian and Russian. In Slovenian such words are the prepositions *v*, *k/h*, *s/z*, consisting of a single consonant (cf. Antić et al., 2006a, 124f.); in Russian these are again prepositions *к*, *с*, *в* and particles *ж* and *б* (cf. Girzig, 1997; Kelih, 2007). In all Slavic languages, these prepositions follow a general trend: they had originally one syllable and have been over time shortened to zero-syllable words.

Subsequent to the operational definition of the linguistic unit word and an adequate choice of the measurement unit, decision with regard to the zero-syllable words is made. These words are not stressed, and should thus be proclitically or enclitically attached to the succeeding or preceding word, respectively. In other words,

[11] The Sorbian languages are the native languages of the Sorbs, a Slavic minority in Eastern Germany. There are two standard varieties of Sorbian: Upper Sorbian which is mainly used in Saxony and Lower Sorbian spoken in Brandenburg by only 28% of all 70.000 Sorbian speakers. The area where the two languages are spoken is known as Lusatia.

[12] For the details on the Göttingen project see `http://wwwuser.gwdg.de/~kbest`

the analysis should be performed without zero-syllable words, hence we consider the orthographic-phonetic definition. The crucial arguments in favor of this decision are: (i) the orthographic-phonetic criterion offers clear-cut principles making the automatic text processing very simple, (ii) this criterion is used by many researchers, thus allowing for comparability of the results and (iii) the proportion of zero-syllable words in Russian and Slovenian texts turned out to be relatively small compared to other x-syllable words. Even many poetic texts do not contain any zero-syllable words at all (cf. Antić et al., 2006a, p. 129).

1.6 Technical Preparation of Texts

At the very beginning the first important step is to prepare everything necessary for a systematic and automatic investigation of the word length frequencies in Slavic languages we are dealing with. Texts have been collected from several available sources such as internet, cd-rom or by very time-consuming text scanning. This has produced differently coded texts. In order to be able to work with these texts, they had to be converted into a unique format (Unicode Standard), enabling as such compatibility with special software. Each text has been saved as a file, as well as a meta-data file containing the information not only about the language, author and title, but also about processing status and computed statistical characteristics and has been stored in a SQL text data base[13]. Consequently, this enables both a possibility to obtain individual texts and extract subgroups of texts according to some criteria.

Special software, developed in the programming language PERL, is used to automatically compute statistical measures such as mean, variance, higher moments, distribution parameters, etc. as well as to determine word lengths and frequency with which words of a given length occur in an individual text. Further informations can be found in Kelih et al. (2003).

1.7 Principles of Word Length Counting

The automatic processing and quantitative analysis of texts consider primarily appropriate choice of the definition of a word and determination of the measurement unit of its length given in Section 1.4. In the process of automatically counting word lengths each text is further submitted to particular tagging procedures, developed in the framework of the Graz project on Quantitative Text Analysis (QuanTA)[14].

[13] The texts are stored in the QuanTA database. For both Slovenian and Russian, over 1000 texts from different authors and text types originating from 19th and 20th century have been collected.

[14] The Graz research project "Word Length (Frequencies) in Slavic Texts" was financially supported by the Austrian Fund for Scientific Research (FWF, P-15485). Further details can be found at `http://www-gewi.uni-graz.at/quanta`.

Similar criteria have been applied in word length studies on other Slavic languages (cf. Wilson, 2006; Rottmann, 2006). The idea is that only the so called "running text" should be analyzed, thus the certain text components that possibly differ from the word length structure of the "running text" should be excluded from the automatic word lengths analysis. The advantage of this approach is the possibility to unify processing of a huge number of texts. To satisfy this criteria it was necessary to tag e.g. (i) titles, subtitles and chapter titles, (ii) comments not written by the author himself, such as foreword, editorial comment or footnote of the publisher, (iii) book number, chapter number and verse number, (iv) text classification given by the author himself, e.g. "ballad" as a subtitle of a ballad, (v) epilogue or prologue, etc.[15] In the process of text unification a couple of additional textual modifications arise from the definition of *word* chosen above (cf. Antić et al., 2006a):

(a) *revision of acronyms*: being realized as a sequence of capitals from the words' initial letters or as letters separated by punctuation marks, acronyms are transformed into corresponding unabbreviated pronunciation textual form. In the examples below, the acronyms are counted as words with two or three syllables.

SMS	(Slovenska mladinska stranka)	→	EsEmEs
SDS	(Socialdemokratska stranka Slovenije)	→	EsDeEs
NK	(Nogometni klub)	→	EnKa
JLA	(Jugoslovanska ljudska armada)	→	JeLeA

(b) *revision of abbreviations*: they are completely transformed into the appropriate full form, e.g.:

c.k.	→	cesarsko–kraljevi
sv.	→	sveti, svetega
g.	→	gospod

(c) *revision of numerals* (especially years), in the form of Arabic or Latin figures (year, date, etc.): numerals are written in their complete form (full words), e.g. year "1907" in the sentence below is counted as three words consisting of seven syllables.

Bilo je leta *1907* → Bilo je leta *tisoč devetsto sedem.*

(d) *revision of hyphenated words* (including hyphenated adjective and noun composites): these are counted as two words, such as "Beneži–Najstati".

[15] For details of tagging procedures see `http://www-gewi.uni-graz.at/quanta/projects/quanta_server/quanta_tags.html`

(e) *revision of words and passages written in foreign languages*: in the case when they appear as single elements in the text they are processed according to their syllabic structure. For instance the name "Wiener Neustadt" occurring in a Slovenian text will be "transliterated" as *Viner Nejstadt* in order to guarantee the underlying Slavic sounded syllabic structure. Furthermore, attention is paid to syllabic and non-syllabic elements which for the two considered languages can differ in function, as for e.g. the letter "Y" in *lorry* → *lori* but *New York* → *Nju Jork*. Longer passages written in a foreign language are completely eliminated.

It should be also noted that independent of these manipulations, the original text structure remains fully recognizable for a researcher and open for further examination. The texts revised in this way are now ready for the concrete statistical analysis.

1.8 Word Length Frequency Distribution

Let $w_1, w_2, \ldots, w_n$ denote different words in a linguistic text of size n, where w_j refers to the j-th word in the text. The set $W = \{w_1, w_2, \ldots, w_n\}$, consisting of all possible words of the given text, corresponds to the sample space. The length of the word $w_j \in W$, denoted by $l(w_j)$ is measured by the number of syllables per word consistent with the principles of automatic text analysis specified in Section 1.7. According to this principles zero-syllable words as parts of the subsequent word do not exist as a separate word class (cf. Section 1.5). Therefore, a certain linguistic text can be seen as a collection of one, two, three, or maximally k syllable words. By counting the number of words of the same length $l(w_j) = i$, $i = 1, \ldots, k$, we observe the frequency f_i with which words of a certain length i appear in a given text. The number of elements f_i that belong to the same frequency class i is called absolute frequency. Relative frequencies, denoted by p_i, are calculated as f_i/n. The total number of words in a given text (text length), n, the absolute frequencies, f_i and the relative frequencies, p_i, are related through $\sum_{i=1}^{k} f_i = n$ and $\sum_{i=1}^{k} p_i = 1$. Collecting the words of the same length into one class leads to the frequency distribution of the word length or frequency distribution of i-syllable words (cf. Table 1.4). In other words, the collection of values i with their associated frequencies f_i, is called frequency distribution of word length occurrence for a given sample. For illustration purposes, consider the first sentence of a Slovenian prose text "Hlapec Jernej in njegova pravica" (I. Cankar, Chapter 1):

> *To povest vam pripovedujem, kakor se je po resnici vršila z vsemi svojimi nekrščanskimi krivicami in z vso svojo veliko žalostjo.*

This sentence contains 19 words, since the zero-syllable word *z* in *z vsemi* and *z vso* is not counted as a separate word. Table 1.3 presents the word length list of this

Table 1.3: The length $l(w_i)$ of the words w_i: "Hlapec Jernej in njegova pravica", Chapter 1, first sentence

i	w_i	$l(w_i)$	i	w_i	$l(w_i)$	i	w_i	$l(w_i)$
1	to	1	8	po	1	15	in	1
2	povest	2	9	resnici	3	16	z vso	1
3	vam	1	10	vršila	3	17	svojo	2
4	pripovedujem	5	11	z vsemi	2	18	veliko	3
5	kakor	2	12	svojimi	3	19	žalostjo	3
6	se	1	13	nekrščanskimi	5			
7	je	1	14	krivicami	4			

sentence. For each word w_i, the length $l(w_i)$ is specified. The words *pripovedujem* and *nekrščanskimi* are the only two five-syllable words. There is only one four-syllable word, five three-syllable words and four two-syllable words. The remaining seven words are one-syllable words. This frequencies are strictly bound to this particular sentence. Obviously, the increase of the sample size from 19 words of the first sentence to 584 words in the whole Chapter 1 of "Hlapec Jernej in njegova pravica" leads to a quite different frequency distribution, as apparent in Table 1.4.

Table 1.4: Frequency distribution of "Hlapec Jernej in njegova pravica" (Chapter 1)

# Syllable per word (i)	absolute frequencies (f_i)	relative frequencies (p_i)
1	259	0.4435
2	195	0.3339
3	92	0.1575
4	35	0.0599
5	3	0.0051

In the textual surrounding a random variable X is a function $X : W \to \mathbb{R}$ which assigns to each word $w_j \in W$ of a particular text one and only one real number $X(w_j) = i$, namely its length. The random variable X denotes hereby the number of syllables per word. The range of X is $\{1, 2, \ldots, k\}$, since no infinitely long words exist in a language (cf. Wimmer et al., 1994, p. 104). As our extensive textual data show, words with more than nine syllables are very rare in Slovenian and Russian. A similar trend can be observed for many other Indo-European languages. As the range of X is finite, the random variable X has a discrete distribution with probability mass function (pmf) given by

$$\pi_i = P(X = i) = P(\{w_j \in W : X(w_j) = i\}) \qquad \text{for all } i, \tag{1.11}$$

where $0 \leq \pi_i \leq 1$ for all i and $\sum_i^k \pi_i = 1$ with $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$.

Statistical distributions serve to model populations and are commonly defined by one or more parameters which determine its functional form. In order to emphasize the dependence of the distribution on the parameter vector $\mathbf{\Theta} = (\theta_1, \theta_2, \ldots, \theta_q)$ from a given parameter space Ω we write $\pi_{i|\Theta} = P_\Theta(X = i)$ instead of $\pi_i = P(X = i)$. In the following chapter special attention is given to the Poisson distribution and its generalizations and we are interested in inferring about $\mathbf{\Theta}$.

Chapter 2

General Approaches

2.1 Restrictions of the Poisson Model

An essential question when modeling count data is how to choose an appropriate probability model to describe the observed values. The simplest and probably the most widely used distribution for analyzing count data is the Poisson distribution. It is primarily used to model the number of events occurring over a particular period of time (or in a given area) with low probability. Therefore, the Poisson distribution is also called the *law of rare events*. The underlying randomness requires independence of these events and constant average number of occurrences in the given time interval (or region of space). Haight (1967) mentioned a multitude of references regarding Poisson applicability in industry, agriculture, biology, medicine, telephony, sociology and demography as well as traffic flow theory. In numerous cited examples the Poisson distribution is applied to the number of defects in material, the number of victims of specific diseases like cholera or malaria, and the number of car accidents in a fixed length of roadway, to mention just some of them. Furthermore, as already noted in Section 1.2, the Poisson distribution and its generalizations proved to be useful for linguistic data regarding word length studies of Slavic languages.

The Poisson distribution depends on a single positive parameter and has probability mass function (pmf) given by

$$\pi_x = P(X = x) = \frac{\theta^x e^{-\theta}}{x!}\,, \quad x = 0, 1, \ldots\,, \quad \theta \in \mathbb{R}^+\,. \tag{2.1}$$

Some examples of the Poisson distribution shown in Figure 2.1 for different values of θ illustrate its flexible form. From its probability generating function (pgf)

$$G_X(t) = \sum_{x=0}^{\infty} t^x P(X = x) = e^{-\theta} \sum_{x=0}^{\infty} \frac{(t\theta)^x}{x!} = e^{\theta(t-1)}\,, \tag{2.2}$$

we can derive simple expressions for the characteristic function (cf), moment generating function (mgf), factorial moment generating function (fmgf), cumulant generating function (cgf) and factorial cumulant generating functions (fcgf); see Table 2.1.

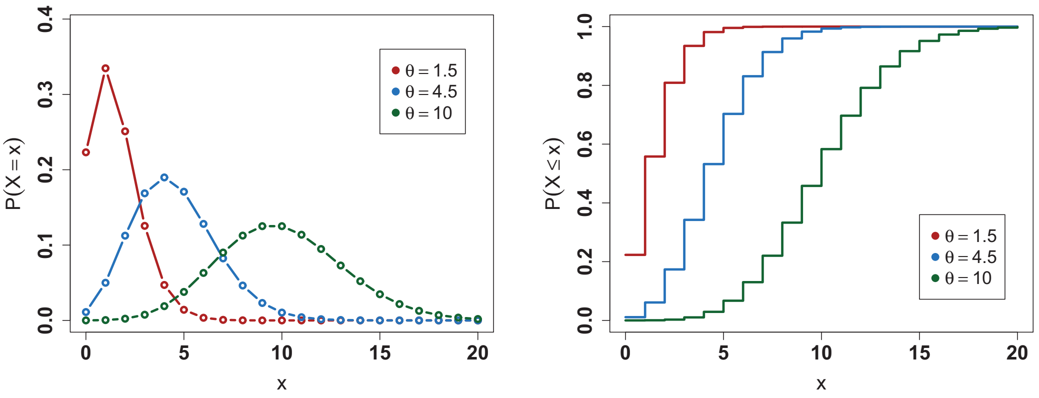


Figure 2.1: Examples of the Poisson probability mass function (left) and its cumulative distribution function (right) for diverse values of parameter θ with mass points at integer values of x only. The connecting lines potentiate only visualization of the functional form.

The mean and the variance are obtained as $\mathrm{E}(X) = \mathrm{var}(X) = \theta$. A short introduction to generating functions and their relations to the theoretical moments is given in Appendix B. Another useful property of the Poisson distribution is that it is *closed under addition*. Closed under addition means that the sum of any two independent count variables coming from the same parametric model also belongs to the same parametric model. Specifically, if X_1 and X_2 are independent with $X_i \sim \mathrm{Poisson}(\theta_i)$ for $i = 1, 2$ then $X_1 + X_2 \sim \mathrm{Poisson}(\theta_1 + \theta_2)$. This result generalizes analogously to the sum of more than two Poisson observations. An important practical consequence of this result is the possibility to analyze grouped data with equivalent results. For further interesting properties of the Poisson model look at the book of Haight (1967) and Johnson et al. (1992).

Table 2.1: Generating functions of $X \sim \mathrm{Poisson}(\theta)$

pgf	$G_X(t) = e^{\theta(t-1)}$	fmgf	$G_X(1+t) = e^{\theta t}$
cf	$\varphi_X(t) = e^{\theta(e^{it}-1)}$	cgf	$K_X(t) = \theta(e^t - 1)$
mgf	$M_X(t) = e^{\theta(e^t-1)}$	fcgf	$\ln G_X(1+t) = \theta t$

The equality of mean and variance, known as *equidispersion*, is an important characteristic of the Poisson distribution. However, this requirement is sometimes too restrictive for the count data. In many practical situations empirical data sets exhibit departures from equidispersion which can be either *overdispersion* (the variance of the count variable exceeds its mean) or *underdispersion* (the variance is smaller than the mean) with respect to the Poisson model and thus cannot be considered to be fitted by the Poisson distribution. A closer look at the distributions of word length in Slovenian and Russian texts shows that texts within one single language, but also between languages, significantly differ with regard to their word lengths. However, as the contribution of Kelih et al. (2005) emphasizes, the word length is rather a characteristic of a genre than a helpful variable to describe an author's individual style. Some typical frequency distributions of word length for four different text types (private letters, journalistic texts, poems and prose texts) are displayed in Figure 2.2. It can be observed that the structure of private letter and prose text do only differ marginally. The most frequent[1] are the one-syllable words, however, even five- and six-syllable words can occur. Compared to this, only very short words are characteristic for the texts from the group of poems. The longest words seem to be typical for journalistic texts, although the frequency of two- and three-syllable words tend to be almost the same for such kinds of texts.

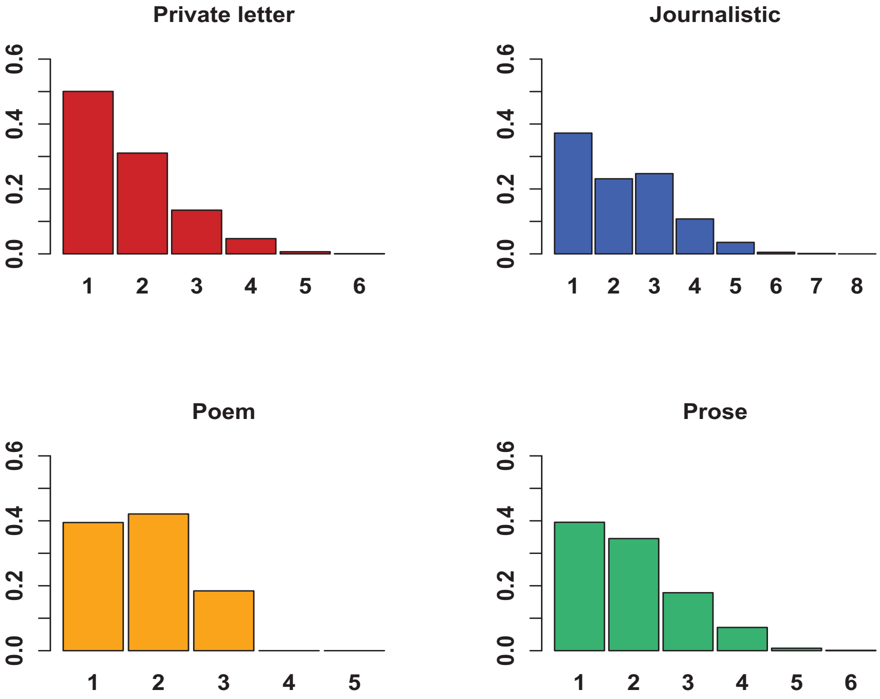


Figure 2.2: Examples of word length frequency distributions for four different text types

To find out whether the count data come from the Poisson distribution we can apply the *index of dispersion* of a count variable X being defined as the variance to mean ratio (cf. Grotjahn, 1982; Puig and Valero, 2006). Due to the fact that texts under study contain no zero-syllable words 1-displaced and size-biased versions of the proposed models became relevant for our further research (for detailed explanation see Section 2.3). For this reason, we propose using the following index of dispersion

$$\delta = \frac{\operatorname{var}(X)}{\mathrm{E}(X) - 1} \tag{2.3}$$

[1] There is a relationship between word frequency and word length saying: "the more frequent a particular word is, the shorter it is". It means that the word length is conditioned by the word frequency, rather than the frequency by the length (cf. Grotjahn, 1982, p. 45).

and estimate it by the empirical value $d = s^2/(\bar{x} - 1)$. Since the 1-displaced Poisson distribution has $\delta = 1$, this model provides an adequate fit only for empirical samples with $d \approx 1$, i.e. for count data where the sample mean diminished by one is near to the sample variance. Therefore, δ can be used as a measure for detecting departures from the 1-displaced Poisson model. Notice that $\delta > 1$ holds for the 1-displaced negative binomial models. Although this family of distributions may arise from various random scenarios, they are likely to be adequate only for empirical samples with overdispersion. However, 1-displaced Dacey-Poisson distributions having $\delta \leq 1$ offer potential solutions in case of underdispersion. As to the problem of Poisson overdispersion various Poisson modifications have been suggested in the literature using continuous discrete mixtures of Poisson distributions, an approach we consider in Section 2.2.1. The same behavior is also typical for Poisson-stopped-sum distributions. Section 2.2.2 gives explanation how to construct such generalized Poisson distributions. Apart from that, the real-life data quite frequently manifest overdispersion through excess of zeros. The zero-inflated distributions, introduced in Section 2.2.1, can be useful in such situations. Moreover, to accomplish discrete distributions that allow for over- and underdispersion a further possibility is to modify Poisson distributions by weighting. Section 2.3.2 gives more details on this topic.

In the following chapters our aim is to find a general theoretical model for word length frequency distributions that covers the whole d range. If possible, the model should be unique for all texts under study and should have at most two parameters.

2.2 Constructions of Distributions

In this section we discuss three different approaches for constructing new distributions that lead to *mixed* or *compound distributions*, *generalized* or *stopped-sum distributions* and finally *misrecorded distributions*. Special attention is paid to the generalization of the Poisson. It is also demonstrated that compounding and generalizing are related concepts. However, the choice of the most adequate distribution for the observations at hand must be justified later based on the sample data.

2.2.1 Mixtures of Discrete Distributions

We distinguish two types of mixtures of discrete distributions, namely finite mixtures and countable or alternatively continuous mixtures (see e.g. Johnson et al., 1992).

Let $\{F_j(x)\}$ represent a set of various proper cumulative distribution functions (cdf's) of different random variables X_j, $j = 1, 2, \ldots, k$. The weighted average

$$F(x) = \sum_{j=1}^{k} \omega_j F_j(x)\,, \tag{2.4}$$

with weights $\omega_j \geq 0$, $\sum_{j=1}^{k} \omega_j = 1$ and finite number of components k, defines the k-*component finite mixture distribution.* If the components are discrete distributions, then the resulting mixture distribution has the probability mass function (pmf) and the probability generating function (pgf) defined respectively by

$$\pi_x = P(X = x) = \sum_{j=1}^{k} \omega_j P(X_j = x) \quad \text{and} \quad G(t) = \sum_{j=1}^{k} \omega_j G_j(t)\,, \tag{2.5}$$

where $G_j(t)$ denotes the pgf of a random variable X_j. The component distributions do not have to be defined over the same sample space. Rather, the sample space of the mixture distribution (2.4) is the union of the sample spaces of the individual components of the mixture. The component distributions are very often of the same type and differ possibly in parameters, although this need not be the case. The k-component finite Poisson mixture with pmf

$$P(X = x) = \sum_{j=1}^{k} \omega_j \frac{e^{-\theta_j}\theta_j^x}{x!} \tag{2.6}$$

is one such example. In most of the cases the number of components k is quite small. The generalized Poisson distribution, considered in Chapter 3, is such a sum of weighted Poisson probabilities with an identical parameter. Figure 2.3 shows three examples of two-component Poisson mixtures, demonstrating that they can be both unimodal as well as bimodal. Finite mixtures of discrete distributions, in particular Poisson and binomial mixtures are described in some detail in Johnson et al. (1992).

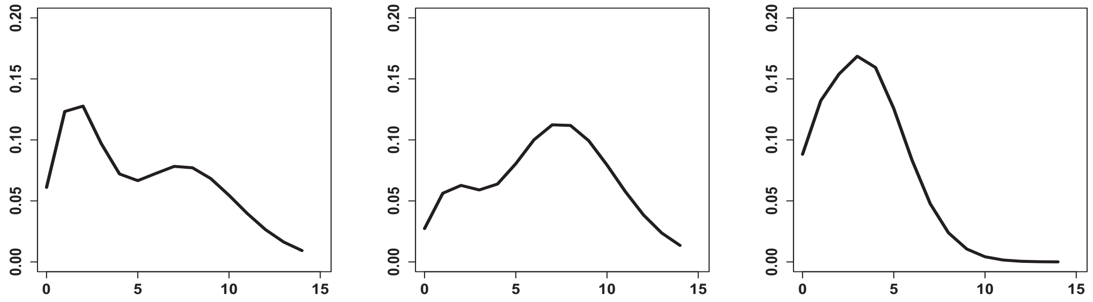


Figure 2.3: Examples of various two Poisson mixtures (left: $\theta_1 = 2\,, \theta_2 = 8\,, \omega_1 = 0.45$; middle: $\theta_1 = 2\,, \theta_2 = 8\,, \omega_1 = 0.2$; right: $\theta_1 = 1\,, \theta_2 = 4\,, \omega_1 = 0.2$)

Zero-inflated distributions are very special two-component finite mixtures (cf. Cameron and Trivedi, 1998; Winkelmann, 2000). Here, the original distribution with pmf π_x^{parent} and support at $x = 0, 1, \ldots$, is combined with the degenerate distribution whose all probability mass is cumulated at zero point, i.e. whose pmf is

$$\pi_x^{deg} = \begin{cases} 1, & \text{for } x = 0\,, \\ 0, & \text{otherwise}\,. \end{cases} \tag{2.7}$$

Hence, based on the equation (2.5) the resulting mixture has the pmf defined by

$$P(X = x) = (1 - \omega)\pi_x^{deg} + \omega\pi_x^{parent}\,, \;\; x = 0, 1, \ldots \tag{2.8}$$

where $0 \leq \omega \leq 1$, or more precisely as

$$P(X = x) = \begin{cases} (1 - \omega) + \omega\pi_0^{parent}\,, & x = 0\,, \\ \omega\pi_x^{parent}\,, & x = 1, 2, \ldots \end{cases}$$

Other synonyms mentioned in the literature for this type of mixture are a *zero-modified distribution* or a *distribution with added zeros* (cf. Johnson et al., 1992). A nice feature of zero-inflated distributions is that their pgf and moments are easily obtained from those of the original distribution. Since the pgf of π_x^{deg} equals unity, the pgf of the modified distribution (2.8) is

$$G_X(t) = (1 - \omega) + \omega H(t)\,, \tag{2.9}$$

where $H(t)$ identifies the original pgf. Similarly, if $\mu_{(k)}^{parent}$ are the factorial moments of the original distribution, then the k-th factorial moment of the mixture (2.8) may be derived using the general formula

$$\mu_{(k)} = \omega\mu_{(k)}^{parent}\,, \;\; k = 1, 2, \ldots \tag{2.10}$$

Also, using equation (B.9), Appendix B, it can easily be proved that the same relation holds for the raw moments. The central moments can then be derived using the standard relations between the raw and central moments given in Appendix B. Hence, the variance is given by

$$\mu_2 = \omega\mu_2^{parent} + \omega(1 - \omega)(\mu^{parent})^2\,. \tag{2.11}$$

Choosing the original distribution to be a Poisson, yields the *zero-modified Poisson distribution*, discussed in more detail in Chapter 4. An important feature of this distribution is its ability to model over- and even underdispersion for $\omega > 1$. Moreover, for $\omega = 1$ it reduces to the standard Poisson distribution, the equidispersed case.

Nevertheless, the mixture distribution can also arise when the distribution function of a random variable X depends on some parameter Θ which is itself a random variable distributed according to a different probability law. The conditional distribution of $X|\Theta$ and the marginal distribution of Θ are known and what we are interested in is the marginal distribution of X. If Θ is defined on the nonnegative integers, the resulting mixture, called *countable mixture*, is obtained by summing the joint distribution of X and Θ over all values θ. Hence, the marginal distribution of X is given by

$$P(X = x) = \sum_{\theta \in \Omega} P(X = x, \Theta = \theta) = \sum_{\theta \in \Omega} P(X = x|\theta)P(\Theta = \theta)\,, \tag{2.12}$$

where Ω denotes the parameter space. As an illustration, consider the mixture of binomial and Poisson where $X|N \sim \text{Binomial}(N, p)$ is combined with $N \sim \text{Poisson}(\theta)$. The summation of $P(X = x, N = n)$ over n yields then

$$\begin{aligned} P(X = x) &= \sum_{n=x}^{\infty} \binom{n}{x} p^x q^{n-x} \frac{\theta^n e^{-\theta}}{n!} = \frac{e^{-\theta}(p\theta)^x}{x!} \sum_{n=x}^{\infty} \frac{(q\theta)^{n-x}}{(n-x)!} \\ &= \frac{e^{-\theta}(p\theta)^x e^{q\theta}}{x!} = \frac{e^{-p\theta}(p\theta)^x}{x!}\,, \end{aligned} \tag{2.13}$$

since $q = 1 - p$, which is the pmf of the Poisson distribution with parameter $p\theta$. Notice that the parameter n has been summed out, hence does not appear in the mixture distribution. Another well-known example of countable mixture is the Neyman Type A distribution which results as a Poisson mixture of Poisson distributions (cf. Neyman, 1939; Feller, 1943; Altmann and Hammerl, 1989).

A *continuous mixture* occurs when parameter Θ has a continuous distribution with probability density function $f(\theta)$. Then the pmf of the mixture is obtained by integrating the joint distribution of X and Θ over the mixing parameter Θ, i.e.

$$P(X = x) = \int_{\theta \in \Omega} P(X = x, \Theta = \theta)\, d\theta = \int_{\theta \in \Omega} P(X = x|\theta) f(\theta)\, d\theta\,. \tag{2.14}$$

The support of the mixture is the same as the support of the initial distribution. Again, parameter θ does not exist in the resulting mixture, since it has been integrated out. The most famous example of a continuous mixture for count data is the negative binomial distribution obtained as a gamma mixture of Poisson distributions. Assume that $X|\Theta \sim \text{Poisson}(\Theta)$ where Θ is gamma distributed with density function

$$f(\theta) = \frac{1}{\beta^\alpha \Gamma(\alpha)}\, \theta^{\alpha-1} e^{-\theta/\beta}\,, \quad \theta, \alpha, \beta > 0\,. \tag{2.15}$$

Then

$$\begin{aligned} P(X = x) &= \int_0^\infty \frac{\theta^x e^{-\theta}}{x!}\, \frac{1}{\beta^\alpha \Gamma(\alpha)}\, \theta^{\alpha-1} e^{-\theta/\beta}\, d\theta \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha) x!} \int_0^\infty \theta^{\alpha+x-1} e^{-\theta(1+1/\beta)}\, d\theta\,, \end{aligned}$$

which by taking $t = -\theta(1+\beta)/\beta$ and using Euler's definition of the gamma function yields

$$P(X = x) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha) x!} \left(\frac{\beta}{\beta+1}\right)^x \left(\frac{1}{\beta+1}\right)^\alpha = \binom{\alpha + x - 1}{x} \left(\frac{1}{\beta+1}\right)^\alpha \left(\frac{\beta}{\beta+1}\right)^x.$$

Hence, X has a negative binomial distribution with parameters α and $1/(\beta+1)$ (cf. Wimmer and Altmann, 1999, p. 449). The hyper-Poisson model, treated in more detail in Chapter 5, is another example of a mixture of Poisson distributions using the truncated Pearson type III mixing distribution. The distribution function of X and Θ may also depend on additional parameters and the procedure can be repeated. For instance, assuming $X|(N,P) \sim \text{Binomial}(N,P)$ where both binomial parameters are randomized as $N \sim \text{Poisson}(\lambda)$ and $P \sim \text{Beta}(\alpha,\beta)$ results in the Beta-Poisson distribution with parameters λ, α and β, which has found its successful application in problems of underreporting (cf. Neubauer and Djuraš, 2009).

Gurland (1957) called mixture distributions of type (2.12) and (2.14) *compound distributions*[2] and introduced the symbolic notation

$$\mathcal{F}_1 \bigwedge_{\Theta} \mathcal{F}_2\,, \tag{2.16}$$

where $\mathcal{F}_1$ represents the original conditional distribution of $X|\Theta$ and $\mathcal{F}_2$ the mixing distribution, which is that of Θ. If the distribution $\mathcal{F}_1$ depends on several parameters $\theta_1, \theta_2, \ldots, \theta_m$, then only the parameter involved in the compounding operation will be specified under the $\bigwedge$ sign. Instead of symbols $\mathcal{F}_1$ and $\mathcal{F}_2$, it is also convenient to write the names of the distributions with corresponding parameters given in parentheses. If the support of Θ is discrete, then by utilizing relation (2.12) the pgf of X results in

Table 2.2: Examples of some compound Poisson distributions with their pgf's

Distribution of X	Derivation	$G_X(t)$
Negative binomial$(\alpha, 1/(\beta+1))$	$\text{Poisson}(\Theta) \bigwedge_{\Theta} \text{Gamma}(\alpha,\beta)$	$1/(1-\beta(t-1))^{\alpha}$
Pascal-Poisson(θ,k,p)	$\text{Poisson}(\theta L) \bigwedge_{L} \text{Negative binomial}(k,p)$	$p^k/(1-qe^{\theta(t-1)})^k$
Beta-Poisson(α,β)	$\text{Poisson}(\Theta) \bigwedge_{\Theta} \text{Beta}(\alpha,\beta)$	${}_1F_1[\alpha;\alpha+\beta;t-1]$
Binomial-Poisson(θ,n,p)	$\text{Poisson}(\theta L) \bigwedge_{L} \text{Binomial}(n,p)$	$(q+pe^{\theta(t-1)})^n$
Hyper-Poisson(θ,λ)	$\text{Poisson}(\theta L) \bigwedge_{L}$ truncated Pearson type III(θ,λ)	$\dfrac{{}_1F_1[1;\lambda;\theta t]}{{}_1F_1[1;\lambda;\theta]}$

$$G_X(t) = \sum_{x=0}^{\infty} t^x P(X=x) = \sum_{x=0}^{\infty} t^x \sum_{\theta \in \Omega} P(X=x|\theta)P(\Theta=\theta)\,,$$

which after the change of the order of summation becomes

$$G_X(t) = \sum_{\theta \in \Omega} P(\Theta=\theta) \sum_{x=0}^{\infty} t^x P(X=x|\theta) = \sum_{\theta \in \Omega} G_X(t|\theta)P(\Theta=\theta)\,,$$

whereas for continuous support of parameter Θ, an integral replaces the summation. Consequentaly, the pgf of X is given by

$$G_X(t) = \mathrm{E}_\theta(G_X(t|\theta))\,. \tag{2.17}$$

[2] Altmann and Zörnig (1992) denote it by *Zusammensetzung* of distributions.

The mean and variance of the marginal distribution of X under mixing can easily be calculated by (cf. Casella and Berger, 2002, 164ff.)

$$\mathrm{E}(X) = \mathrm{E}_\theta(\mathrm{E}(X|\theta))\,, \quad \mathrm{var}(X) = \mathrm{E}_\theta(\mathrm{var}(X|\theta)) + \mathrm{var}_\theta(\mathrm{E}(X|\theta))\,, \tag{2.18}$$

provided that the expectations exist. However, specifying the conditional distribution of $X|\Theta$ to be Poisson and using its equidispersion property we have

$$\mathrm{E}_\theta(\mathrm{var}(X|\theta)) = \mathrm{E}_\theta(\mathrm{E}(X|\theta)) = \mathrm{E}(X)\,,$$

which after substitution for variance in (2.18) results in overdispersion at the marginal level of X, no matter what the distribution of Θ is and due to the fact that

$$\mathrm{var}(X) = \mathrm{E}(X) + \mathrm{var}_\theta(\mathrm{E}(X|\theta)) > \mathrm{E}(X)\,. \tag{2.19}$$

This special type of mixture where $X|\Theta$ has a Poisson distribution is called the *compound Poisson distribution* by Feller (1943) and Maceda (1948), although the notation *mixture of Poisson distributions* used above is even more common. The pgf of X is obtained by substituting $G_X(t|\theta) = e^{\theta(t-1)}$ in (2.17), hence it becomes

$$G_X(t) = \begin{cases} \sum\limits_{\theta\in\Omega} e^{\theta(t-1)} P(\Theta = \theta), & \text{if } \Theta \text{ dicrete}\,, \\ \int\limits_{\theta\in\Omega} e^{\theta(t-1)} f(\theta), & \text{if } \Theta \text{ continuous}\,, \end{cases} \tag{2.20}$$

thereby the random variable Θ can take only positive real values. The resulting mixture is necessarily a discrete distribution with jumps at the nonnegative integers (cf. Teicher, 1960). This is also apparent from the examples in Table 2.2. Furthermore, it is interesting to note that the k-th factorial moment of a mixture of Poisson distributions is obtained from (2.20) as

$$\mu_{(k)} = G_X^{(k)}(1) = \mathrm{E}(\Theta^k)\,, \quad k = 1, 2, \ldots \tag{2.21}$$

thus being equal to the k-th raw moment of the mixing distribution. Notice also, that the factorial moment generating function (fmgf) of a mixture of Poisson distributions $G_X(1+t)$ is equal to the moment generating function (mgf) of the mixing distribution $M_\Theta(t)$ (see Appendix B for definitions). Another important feature, proved by Feller (1943, p. 394), states that the convolution of two mixtures of Poisson distributions is again a mixture of Poisson distributions with a mixing distribution that is the convolution of the two given mixing distributions. A quite large number of various continuous and countable mixtures of Poisson distributions can be found in Johnson et al. (1992) and Altmann and Zörnig (1992). Among them are also examples given in Table[3] 2.2 and Table 2.3, in the next section.

[3] In Table 2.2 above ${}_1F_1$ denotes the confluent hypergeometric function. For further details see Chapter 5.

2.2.2 Generalized Distributions

The distributions considered in this section are based on the random summation of independent and identically distributed random variables.
Let $X_1, X_2, \ldots, X_N$ be mutually independent random variables with common distribution $P(X = j) = \pi_j$ and pgf $G_X(t) = \sum_{i=0}^{\infty} t^i \pi_i$. Consider a random variable

$$S_N = X_1 + X_2 + \ldots + X_N = \sum_{i=1}^{N} X_i \,, \tag{2.22}$$

where the number of components N is a random variable independent of X_i, for each $i = 1, \ldots, N$. Let $P(N = n) = g_n$ be the probability function of N and $G_N(t) = \sum_{n=0}^{\infty} t^n g_n$ its pgf. For a fixed n the distribution of $X_1 + X_2 + \ldots + X_n$ is given by the n-fold convolution[4] of X with itself and hence equals $\pi_j * \pi_j * \ldots * \pi_j = \pi_j^{n*}$. Therefore, the distribution of S_N is obtained by the conditional probabilities as

$$P(S_N = j) = \sum_{n=0}^{\infty} P(S_N = j | N = n) P(N = n) = \sum_{n=0}^{\infty} \pi_j^{n*} g_n \,. \tag{2.23}$$

However, Feller (1968, p. 287) showed that it is much easier to calculate the pgf of S_N instead of its distribution (2.23) using the pgf's of X and N. Since the pgf of the sum of independent random variables is equal to the product of the respective pgf's (see relation (B.18), Appendix B), we obviously have $G_{S_N|N}(t) = \mathrm{E}(t^{S_N} | N = n) = G_X^n(t)$. Consequently, after a permitted change of the order of summation, we have

$$G_{S_N}(t) = \sum_{j=0}^{\infty} t^j P(S_N = j) = \sum_{n=0}^{\infty} g_n \sum_{j=0}^{\infty} t^j \pi_j^{n*} = \sum_{n=0}^{\infty} g_n G_X^n(t) \,.$$

The last expression represents the pgf of the random variable N, where t is replaced by the whole pgf of X, thus

$$G_{S_N}(t) = G_N(G_X(t)) \,. \tag{2.24}$$

Notice that this identity can also be derived by use of relation (2.18) since

$$G_{S_N}(t) = \mathrm{E}(t^{S_N}) = \mathrm{E}_N(\mathrm{E}(t^{S_N} | N)) = \mathrm{E}_N(\mathrm{E}^n(t^X)) = \mathrm{E}_N(G_X^n(t)) = G_N(G_X(t)) \,.$$

Substituting above any particular distribution function for N we can get a certain form of $G_{S_N}(t)$, so that (2.24) determines a whole family of distributions (see the class of generalized Poisson distributions below).

Denote further by $\varphi_N(t) = G_N(e^{it})$ and $\varphi_X(t) = G_X(e^{it})$ the characteristic functions (cf's) of random variables N and X, respectively. Then we get the cf of the

[4] German writers prefer the notation *Faltung* instead of the term *convolution*. Particulary important is the convolution of identical distributions, denoted by $*$. The pmf of the convolution of two identical distributions is $\pi_x * \pi_x = \pi_x^{2*}$.

random sum S_N by applying (2.24) to $\varphi_{S_N}(t) = G_{S_N}(e^{it})$ and we have (cf. Johnson et al., 1992, p. 345)

$$\varphi_{S_N}(t) = G_N(G_X(e^{it})) = G_N(\varphi_X(t)) = G_N(e^{i(-i\ln\varphi_X(t))}) = \varphi_N(-i\ln\varphi_X(t))\,. \quad (2.25)$$

This model was named *contagion* by Neyman (1939) who developed a new family of contagious distributions of types A, B and C that provide a reasonably good fit for biological phenomena, especially in the population of insects and bacteria. Later on, concentrating primary on the Poisson context some authors, such as Satterthwaite (1942), Feller (1943) and Maceda (1948), have chosen to use the term *generalization* instead to describe such a process. Gurland (1957) maintained the former notation and introduced the symbolic representation of these distributions, given by Definition 2.2 below, in order to facilitate their application. Let us first define the term "equivalent distributions" (cf. Gurland, 1957, p. 266).

Definition 2.1 *Let the random variables Y_1 and Y_2 have cdf's $F_1(y|\alpha)$ and $F_2(y|\beta)$, respectively. If for each α exists some β and for each β exists some α such that $F_1(y|\alpha) = F_2(y|\beta)$ whatever y is, the random variables Y_1 and Y_2 are said to be equivalent. Symbolically we write $Y_1 \sim Y_2$.*

Now the distribution of a random variable S_N is given by the following definition.

Definition 2.2 *Let the random variables N and X have distributions $\mathcal{F}_1$ and $\mathcal{F}_2$ with pgf's $G_N(t)$ and $G_X(t)$, respectively. Then a distribution with pgf of the form $G_N(G_X(t))$ is called a generalized $\mathcal{F}_1$ distribution, or an $\mathcal{F}_1$ distribution generalized by the generalizing $\mathcal{F}_2$ distribution. Symbolically it is represented by*

$$S_N \sim \mathcal{F}_1 \bigvee \mathcal{F}_2\,. \quad (2.26)$$

Thereby, as pointed out by Gurland (1957), distributions $\mathcal{F}_1$ and $\mathcal{F}_2$ need not necessarily to be discrete. Generalized distributions are often referred to as *stopped-sum distributions*, meaning that the number of observations from the distribution $\mathcal{F}_2$ which is to be summed is determined (or stopped) by an observation having the distribution $\mathcal{F}_1$ (cf. Johnson et al., 1992, 343ff.).

The main properties of the generalized distributions are easily obtained using their pgf's. Katti (1966) showed that the factorial moments of distribution (2.26) can be expressed in terms of their component distributions $\mathcal{F}_1$ and $\mathcal{F}_2$. Denote by ${}_1\mu_{(i)}$ and ${}_2\mu_{(i)}$ the factorial moments of distributions $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Then the factorial moments of $\mathcal{F}_1 \bigvee \mathcal{F}_2$ distribution are obtained from the derivatives of $G_{S_N}(t) = G_N(G_X(t))$ at $t = 1$. The first three of them are given by

$$\begin{aligned} &\mu_{(1)} = {}_1\mu_{(1)2}\mu_{(1)}\,, \quad \mu_{(2)} = {}_1\mu_{(2)2}\mu^2_{(1)} + {}_1\mu_{(1)2}\mu_{(2)}\,, \\ &\mu_{(3)} = {}_1\mu_{(3)2}\mu^3_{(1)} + 3{}_1\mu_{(2)2}\mu_{(2)2}\mu_{(1)} + {}_1\mu_{(1)2}\mu_{(3)}\,. \end{aligned} \quad (2.27)$$

Moreover, single probabilities of the generalized distribution (2.26) can be computed by using its component probabilities $g_i = P(N = i)$ and $\pi_i = P(X = i)$. Denote by

$$G_{S_N}(t) = \sum_{i=0}^{\infty} t^i P(S_N = i)\,,\;\; G_N(t) = \sum_{i=0}^{\infty} t^i g_i\,,\;\; G_X(t) = \sum_{i=0}^{\infty} t^i \pi_i\,, \tag{2.28}$$

then using the relation (B.2), Appendix B, we obtain

$$\begin{aligned} P(S_N = 0) &= G_N(G_X(0)) = G_N(\pi_0) = \sum_{i=0}^{\infty} \pi_0^i g_i\,, \\ P(S_N = 1) &= G_N'(G_X(0))\,G_X'(0) = G_N'(\pi_0)\pi_1 = \sum_{i=0}^{\infty} i\pi_0^{i-1} g_i \pi_1\,, \\ P(S_N = 2) &= \frac{1}{2}\,G_N''(\pi_0)\pi_1^2 + G_N'(\pi_0)\pi_2 = \sum_{i=0}^{\infty} g_i \left(\frac{i(i-1)}{2}\,\pi_0^{i-2}\pi_1^2 + i\pi_0^{i-1}\pi_2 \right). \end{aligned}$$

However, this calculation becomes quite tedious and complex for higher probabilities. If S_N has a discrete distribution, the above computation is considerably simplified by the use of formula (B.7) from Appendix B, which enables to get probabilities $P(S_N = i)$ through its factorial moments $\mu_{(i)}$, provided that ${}_1\mu_{(i)}$ and ${}_2\mu_{(i)}$ in (2.27) are known.

Among all random sums S_N of particular interest are those where N has a Poisson distribution. If we denote the expectation of N with θ, then S_N obeys the *generalized Poisson distribution* law. By formula (2.24) its pgf is

$$G_{S_N}(t) = e^{\theta(G_X(t)-1)} = e^{-\theta+\theta G_X(t)}\,, \tag{2.29}$$

where $G_X(t)$ is the pgf of an arbitrary distribution. Interestingly, the k-th convolution of S_N with itself is again a distribution of the same type, only with θ replaced by $k\theta$, since $[G_{S_N}(t)]^k = e^{k\theta(G_X(t)-1)}$. Furthermore, as Gurland (1965, p. 142) indicated, there is a connection between the probabilities of the generalized Poisson distribution and that of X. Based on (B.2), Appendix B, the single probabilities $P_i = P(S_N = i)$ are defined by the derivatives of $G_{S_N}(t)$ at $t = 0$. Thus, derivation of relation (2.29) by setting $t = 0$ yields in first three iterations

$$P_1 = \theta\pi_1 P_0\,,\;\; P_2 = \frac{\theta^2}{2}\,\pi_1^2 P_0 + \theta\pi_2 P_0\,,\;\; P_3 = \frac{\theta^3}{6}\,\pi_1^3 P_0 + 3\theta^2\pi_1\pi_2 P_0 + \theta\pi_3 P_0\,, \ldots \tag{2.30}$$

Consequently, after $k-1$ derivations we obtain the following recurrence formula

$$P_k = \frac{\theta}{k} \sum_{i=0}^{k-1} (i+1)\pi_{i+1} P_{k-1-i}\,. \tag{2.31}$$

Satterthwaite (1942, p. 412) obtained the first four moments of the generalized Poisson distribution by differentiating its cf $\varphi_{S_N}(t) = G_N(\varphi_X(t)) = e^{\theta(\varphi_X(t)-1)}$ which results from equation (2.25). The resulting moments of a generalized Poisson distribution are functions of the raw moments of the underlying distribution due to the fact that

$$\begin{aligned} \mu'_1 &= \theta\, {}_g\mu'_1\,, \qquad \mu_2 = \theta\, {}_g\mu'_2\,, \\ \mu_3 &= \theta\, {}_g\mu'_3\,, \qquad \mu_4 = \theta\, {}_g\mu'_4 + 3\theta^2 ({}_g\mu'_2)^2\,, \end{aligned} \tag{2.32}$$

where ${}_g\mu'_k$ denotes the k-th raw moment of the generalizing distribution of X. Notice that for any X the above variance can be written as $\mu_2 = \theta\,({}_g\mu_2 + {}_g\mu^2) > \theta\, {}_g\mu = \mu$, thus generalized Poisson distributions are characterized by overdispersion.

The generalized Poisson distribution considered in Chapter 6 is obtained by generalizing Poisson through Borel distribution (cf. Wimmer and Altmann, 1999, p. 93). But, we could also assume X_i in (2.22) be independent and identically distributed Bernoulli random variables with probability of success p, for some $p \in (0, 1]$. This process of generalizing the Poisson distribution by the Bernoulli distribution is known as an *independent p-thinning* (cf. Puig and Valero, 2007). Based on equation (2.29), the pgf of S_N results in $G_{S_N}(t) = e^{\theta\,(q+pt-1)} = e^{p\theta(t-1)}$, which is the pgf of the Poisson distribution with parameter $p\theta$, already given in (2.13). Since $S_N|N \sim \text{Binomial}(N, p)$, the resulting random variable S_N can also be understood as a *binomial subsampling* of N. The negative binomial distribution, regarded as a gamma mixture of Poisson distributions, may also be interpreted as a generalized Poisson distribution where the generalization is achieved by logarithmic distribution (cf. Gurland, 1965; Winkelmann, 2000). There are various other distributions that can be considered both as compound and stopped-sum Poisson distributions, one further example being Neyman's two-parametric distribution of Type A (cf. Feller, 1943; Gurland, 1958, 1965). Moreover, Table 2.3 below emphasizes that compounding and generalizing are related construction concepts, as listed distributions are simultaneously of the type (2.16) and (2.26). The following Gurland's (1957) theorem defines the crucial condition under which the same distribution can be generated both as a compound and as a generalized distribution.

Theorem 2.1 *Let X be a random variable with distribution function $F_2(x|\theta N)$ and pgf $G_X(t|\theta N)$ which is dependent on a parameter θ such that*

$$G_X(t|\theta N) = [G_X(t|\theta)]^n\,. \tag{2.33}$$

Let parameter N itself be a random variable with distribution function F_1 and pgf $G_N(t) = \sum_{n=0}^{\infty} t^n g_n$. Then, whatever N is the following relationship holds

$$\mathcal{F}_2(\theta N) \bigwedge_N \mathcal{F}_1 \sim \mathcal{F}_1 \bigvee \mathcal{F}_2(\theta)\,. \tag{2.34}$$

To prove the above theorem we obtain the pgf of the mixture $\mathcal{F}_2(\theta N) \bigwedge \mathcal{F}_1$ using the conditional probabilities

$$G_X(t) = \sum_{x=0}^{\infty} t^x P(X = x) = \sum_{x=0}^{\infty} t^x \sum_{n=0}^{\infty} P(X = x|\theta N) g_n\,,$$

Table 2.3: Examples of some distributions with dual genesis and common pgf's

Distribution of X	Character of derivation	$G_X(t)$
Poisson$(p\theta)$	$\text{Binomial}(N,p) \underset{N}{\bigwedge} \text{Poisson}(\theta)$ $\text{Poisson}(\theta) \bigvee \text{Bernoulli}(p)$	$e^{p\theta(t-1)}$
Poisson-binomial(θ, n, p)	$\text{Binomial}(nL,p) \underset{L}{\bigwedge} \text{Poisson}(\theta)$ $\text{Poisson}(\theta) \bigvee \text{Binomial}(n,p)$	$e^{\theta((q+pt)^n - 1)}$
Binomial-Poisson(θ, n, p)	$\text{Poisson}(\theta L) \underset{L}{\bigwedge} \text{Binomial}(n,p)$ $\text{Binomial}(n,p) \bigvee \text{Poisson}(\theta)$	$\left(q + pe^{\theta(t-1)}\right)^n$
Negative binomial(k,p)	$\text{Poisson}(\Theta) \underset{\Theta}{\bigwedge} \text{Gamma}(k, q/p)$ $\text{Poisson}(-k \ln p) \bigvee \text{Logarithmic}(q)$	$p^k/(1-qt)^k$
Neyman Type A(θ, λ)	$\text{Poisson}(\theta N) \underset{N}{\bigwedge} \text{Poisson}(\lambda)$ $\text{Poisson}(\lambda) \bigvee \text{Poisson}(\theta)$	$e^{\lambda\left(e^{\theta(t-1)} - 1\right)}$
Poisson-Pascal(θ, k, p)	$\text{Negative binomial}(kN,p) \underset{N}{\bigwedge} \text{Poisson}(\theta)$ $\text{Poisson}(\theta) \bigvee \text{Negative binomial}(k,p)$	$e^{\theta\left(p^k(1-qt)^{-k} - 1\right)}$

where $g_n = P(N = n)$ is pmf of N. By interchanging the order of summation and applying Gurland's requirement (2.33) we have

$$G_X(t) = \sum_{n=0}^{\infty} g_n G_X(t|\theta N) = \sum_{n=0}^{\infty} [G_X(t|\theta)]^n \, g_n = G_N(G_X(t|\theta)) \,.$$

The right-hand side of the last expression is the same as (2.24), hence relation (2.34) holds, regardless of distribution $\mathcal{F}_1$. This means that the mixed distribution which satisfies Gurland's precondition has an alternative genesis as the generalized distribution. There is an infinite class of distributions with this property. In particular, discrete mixtures of the following Poisson, binomial and negative binomial distributions with pgf's

$$\begin{array}{ll}
\text{Poisson}(\theta N) & \Rightarrow \; G_X(t|\theta N) = e^{\theta n(t-1)} = \left(e^{\theta(t-1)}\right)^n = [G_X(t|\theta)]^n \,, \\
\text{Binomial}(mN, p) & \Rightarrow \; G_X(t|mN) = (q+pt)^{mn} = ((q+pt)^m)^n = [G_X(t|\theta)]^n \,, \\
\text{Negative binomial}(kN, p) & \Rightarrow \; G_X(t|kN) = \left(\dfrac{p}{1-qt}\right)^{kn} = \left(\left(\dfrac{p}{1-qt}\right)^k\right)^n = [G_X(t|\theta)]^n \,,
\end{array}$$

may also be interpreted as generalized distributions in the sense of Definition 2.2. Some examples are given in Table 2.3. However, not every distribution which is both a compound and generalized distribution can be generated in this manner. The negative binomial distribution given in Table 2.3 is one such example.

2.2.3 Misrecorded Poisson Distributions

These modified distributions consider certain models conforming to the Poisson except for some particular category, which is subject to defective counts. The Cohen-Poisson distribution discussed extensively in Chapter 7 has its application in situations where some, though not necessarily all, values of “ones” are faulty recorded as “zeros” while values of two and more from Poisson are recorded correctly. The data of this type arise very often as a result of misclassification during inspection of the number of defects per unit. An investigator may sometimes fail to see the units which actually have a single defect as being free of defects and hence make errors in data reporting. Cohen (1959, 1991) observes a more general case in which some values of $x = k + 1$ are incorrectly reported as $x = k$ with probability α, where k may be any value of the random variable X and $0 \leq \alpha \leq 1$.

2.3 Effect of Excluded Zero-Syllable Words

In linguistics, very often a need arises to deal with variables that can not take the zero value, such as syllable length measured in number of graphemes, word length measured in number of graphemes, morphemes or even syllables (cf. Section 1.4). Also sentence length measured in number of words can never take a zero value. In some concrete situations even more values at the lower end of the scale can be missing, specifically in the case when the random variable denotes “number of words in the verse”. We discuss here two possibilities that enable dealing with distributions when the zero class is omitted, namely the 1-displaced and the size-biased approach.

2.3.1 1-Displaced Distributions

The 1-displaced distribution corresponds to the distribution of the linear transformation $X^d = X + 1$ of the random variable X. Thereby, if X is a discrete random variable with distribution function $F_X(x)$ then X^d is also a discrete random variable (cf. Casella and Berger, 2002, p. 48).

The first two moments of the random variable X^d can easily be calculated as

$$\mathrm{E}(X^d) = \mathrm{E}(X+1) = \mathrm{E}(X) + 1 \qquad \text{and} \qquad \mathrm{var}(X^d) = \mathrm{var}(X)\,. \tag{2.35}$$

The pgf of X^d is given by

$$G_{X^d}(t) = \mathrm{E}(t^{X^d}) = \mathrm{E}(t^{X+1}) = tG_X(t)\,, \tag{2.36}$$

whereas its mgf equals

$$M_{X^d}(t) = \mathrm{E}(e^{tX^d}) = \mathrm{E}(e^{tX+t}) = e^t M_X(t)\,. \tag{2.37}$$

If all factorial moments $\mu_{(k)}$ of X are known, then by applying formula (2.36) the factorial moments $\mu^d_{(k)}$ of X^d can be evaluated by the following recurrence relation

$$\mu^d_{(k)} = \mu_{(k)} + k\mu_{(k-1)}\,, \quad k = 1, 2, \dots\,, \tag{2.38}$$

where $\mu_{(0)} = 1$ holds. Furthermore, raw and central moments of the random variable X^d can be determined using relations (B.9) and (B.13), Appendix B. Table 2.4 gives several basic discrete distributions and their 1-displaced forms with pmf and first two moments.

Consider X to be Poisson distributed with parameter $\theta \in \mathbb{R}^+$. Then the distribution of the random variable X^d is referred to as *1-displaced Poisson distribution* with pmf defined by

$$\pi^d_x = P(X^d = x) = \frac{e^{-\theta}\theta^{x-1}}{(x-1)!}\,, \quad x = 1, 2, \dots\,, \tag{2.39}$$

where the nonnegative integer valued domain $\mathbb{N}_0$ of X is replaced by the domain $\mathbb{N}$. Since $\mathrm{E}(X^d) = \theta + 1 > \mathrm{var}(X^d) = \theta$ for any $\theta > 0$ this distribution is characterized by underdispersion regarding the standard Poisson distribution. This further implies that linear transformation on the sample space do not generate again the Poisson distribution with a different value of the parameter θ, i.e. the Poisson distribution is not *closed under linear transformation.*

2.3.2 Size-Biased Distributions

Size-biased distributions arise as a special case of a more general class of weighted distributions. Although first introduced by Fisher (1934) as a method to model ascertainment bias, it was Rao (1965) who later formalized this approach in a unifying theory. The concept of weighted distributions has been used mainly in situations where the usual random sampling from the underlying population is not possible due to data having unequal probabilities of being selected in the sample. Even if a random sample can be obtained it may be too difficult or too expensive or possibly less effective to do so. As an example suppose that one is interested in studying the population of criminals in a particular country. Obviously, it would be highly expensive and rather hopeless to sample randomly from the entire population of criminals. An easier way, mentioned by Larose and Dey (1996), would be to investigate the population of criminals already in jail with a distribution which is the weighted version of the original distribution according to the probability of a criminal being caught. Among others, such weighted probability surveys arise as a result of the probability proportional to size, size-biased and line-transect sampling designs (cf. Patil and Rao, 1978; Patil, 2002). Properly defined sampling frames rarely exist for human and wildlife populations which is evident from the study of data on abnormal (albino) children in Rao (1965), the analysis on family size and sex ratio reported in Patil and Rao (1978) or the question of estimating the abundance of a particular animal species (deep-sea red crab and minke whale) in a given region illustrated by Patil (2002). Moreover, weighted distributions are applicable in many other applied fields as agriculture, forestry and ecology (cf. Gove, 2003a), to mention just a few of them. Recorded observations in these populations are biased and will not have the original distribution. In order to make a specification of the probabilities of recorded events it is necessary to adjust the probabilities of actual occurrences of events. The size-biased distributions are applied here in the modelling framework, as a tool to select an appropriate model for the observed data and are not to be connected with the sample selection methods. First, we introduce the concept of weighted distributions.

Suppose that the original (unobserved) pmf of the discrete random variable X is $\pi_{x|\Theta} = P_\Theta(X = x)$ with unknown parameter vector $\mathbf{\Theta}$ from a given parameter space Ω. Assume further that the probability of recording the observation $X = x$ is given by an arbitrary non-negative weight function $w(x, \alpha)$ depending on the value x and possibly on an unknown parameter α. Then the corresponding *weighted distribution* of the random variable X^w has a pmf[5] given by (cf. Patil and Rao, 1978; Patil, 2002)

$$\pi^w_{x|\Theta} = P_\Theta(X^w = x) = \frac{w(x,\alpha)\pi_{x|\Theta}}{\mathrm{E}(w(x,\alpha))}\,, \quad x \in \mathbb{N}\,, \quad \alpha \in \mathbb{R}^+\,, \tag{2.40}$$

where $\mathrm{E}(w(x,\alpha)) = \sum_x w(x,\alpha)\pi_{x|\Theta}$ is the normalizing factor obtained to make the total probability equal to unity. The weighted function does not necessary need to be bounded by $0 \le w(x,\alpha) \le 1$, it may even exceed unity. However, it should be of the form for which $\mathrm{E}(w(x,\alpha))$ exists. Note that $\pi^w_{x|\Theta} = \pi_{x|\Theta}$ holds if and only if $w(x,\alpha)$ is a constant, because then $\mathrm{E}(w(x,\alpha)) = w(x,\alpha)$. Patil and Rao (1978) mentioned that some statistical distributions may be expressed as weighted distributions. For example, a zero-truncated distribution is a special case of distribution (2.40), where $w(0,\alpha) = 0$ and $w(x,\alpha) = 1$ for $x \ge 1$.

An important family of distributions, known as *size-biased distributions*, arises when the weight function in formula (2.40) is taken to be $w(x,\alpha) = x$. Consequently, the pmf of the size-biased version X^*, of a random variable X is of the form

$$\pi^*_{x|\Theta} = P_\Theta(X^* = x) = \frac{x\pi_{x|\Theta}}{\mathrm{E}(X)}\,, \quad x = 1, 2, \ldots\,, \tag{2.41}$$

where the denominator is the mean of the original distribution and $\pi^*_{0|\Theta} = 0$ holds. Some practical examples of such distributions are considered by Rao (1965) and Patil (2002). In forestry, size-biased models are referred to as *length-biased distributions*, since the probability of selecting individuals into the sample is proportional to some linear measure, such as piece length or diameter. Two- and three-parameter size-biased Weibull distributions found wide application as a diameter at breast height distribution models (cf. Gove, 2003a). Their estimating equations for both method of moments and maximum likelihood are presented in Gove (2003b). Akman

[5] Notice that the recorded x is not an observation on the random variable X, but on X^w.

et al. (2007) studied how to identify length-biased samples and provided a simple distribution-free test which only uses the sampled information with the application to the population of grasshoppers. Table 2.4 presents several basic discrete distributions and their size-biased forms with pmf and first two moments. It can be seen that the resulting distribution of X^* is of the same form as the distribution of X for binomial, Poisson and negative binomial distributions with the variable reduced by unity. Patil and Rao (1978) observed the same result for hypergeometric and binomial beta distribution. However, the situation is different for the logarithmic distribution which changes instead to the geometric distribution. Apparently, the size-biased and 1-displaced versions of the Poisson distribution are identical. Comparison of the panels in Figure 2.4 clearly shows the interplay between the Binomial distribution and its 1-displaced and size-biased forms, being plotted as rows, for the sample size of n=10. The columns show the effect of increasing the parameter p by a factor of three.

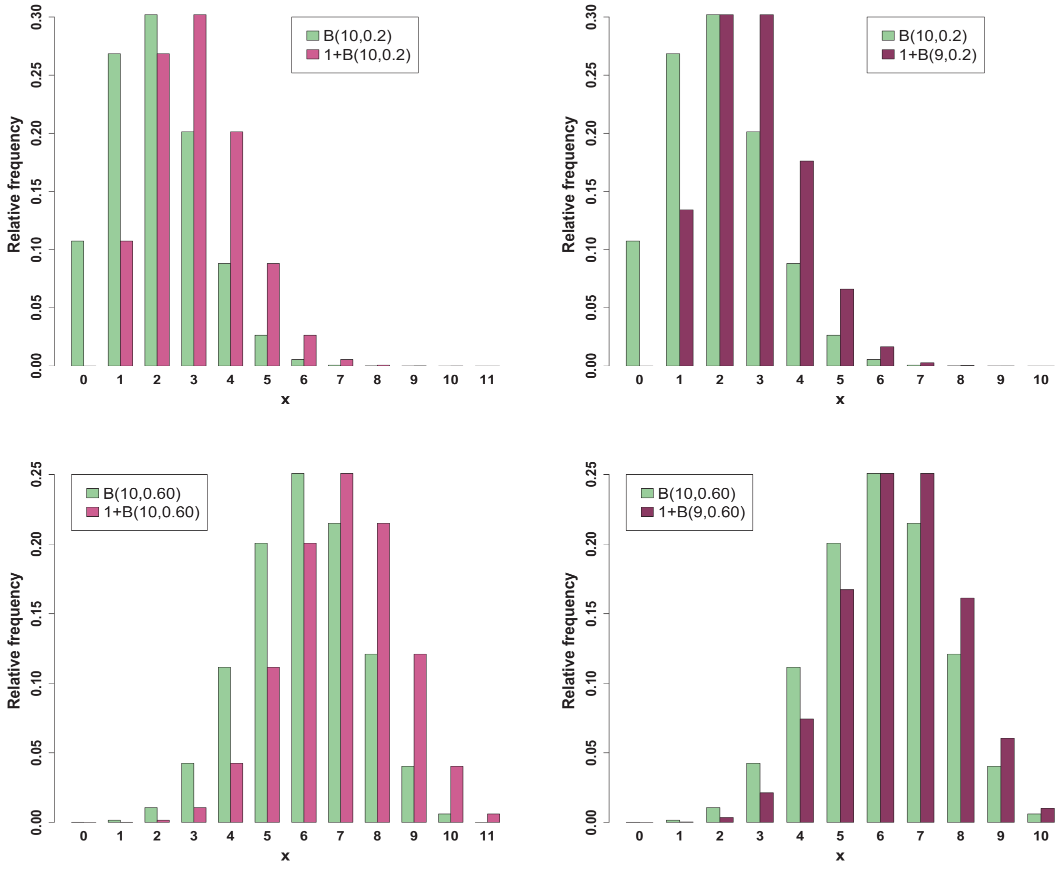


Figure 2.4: Binomial distribution compared to its 1-displaced (left column) and size-biased version (right column) for two different values of $p = 0.2$ and $p = 0.6$ given as rows with $(\mu, \mu^d, \mu^*) = (2, 3, 2.8)$, $s = s^d = 1.26$, $s^* = 1.2$ and $(\mu, \mu^d, \mu^*) = (6, 7, 6.4)$, $s = s^d = 1.55$, $s^* = 1.47$, respectively.

Table 2.4: Some basic discrete distributions with their 1-displaced and size-biased forms

		Random Variable		
Distribution		X	X^d	X^*
Poisson	Notation	$\mathrm{P}(\theta)$	$1{+}\mathrm{P}(\theta)$	$1{+}\mathrm{P}(\theta)$
	Range	$\mathbb{N}_0$	$\mathbb{N}$	$\mathbb{N}$
	pmf	$\dfrac{e^{-\theta}\theta^x}{x!}$	π_{x-1}	π_{x-1}
	E(·)	θ	$1+\theta$	$1+\theta$
	var(·)	θ	θ	θ
Binomial	Notation	$\mathrm{B}(n,p)$	$1{+}\mathrm{B}(n,p)$	$1{+}\mathrm{B}(n-1,p)$
	Range	$\{0,\ldots,n\}$	$\{1,\ldots,n+1\}$	$\{1,\ldots,n\}$
	pmf	$\binom{n}{x}p^xq^{n-x}$	π_{x-1}	$\dfrac{x}{np}\pi_x$
	E(·)	np	$np+1$	$(n-1)p+1$
	var(·)	npq	npq	$(n-1)pq$
Negative Binomial	Notation	$\mathrm{NB}(k,p)$	$1{+}\mathrm{NB}(k,p)$	$1{+}\mathrm{NB}(k+1,p)$
	Range	$\mathbb{N}_0$	$\mathbb{N}$	$\mathbb{N}$
	pmf	$\binom{k+x-1}{x}p^kq^x$	π_{x-1}	$\dfrac{xp}{kq}\pi_x$
	E(·)	$\dfrac{kq}{p}$	$\dfrac{kq+p}{p}$	$\dfrac{kq+1}{p}$
	var(·)	$\dfrac{kq}{p^2}$	$\dfrac{kq}{p^2}$	$\dfrac{(k+1)q}{p^2}$
Logarithmic	Notation	$\mathrm{Log}(p)$	$1{+}\mathrm{Log}(p)$	$\mathrm{G}(p)$
	Range	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$
	pmf	$\dfrac{ap^x}{x}$	π_{x-1}	$p^{x-1}(1-p)$
	E(·)	$\dfrac{ap}{(1-p)}$	$\dfrac{ap}{(1-p)}+1$	$\dfrac{1}{(1-p)}$
	var(·)	$\dfrac{ap(1-ap)}{(1-p)^2}$	$\dfrac{ap(1-ap)}{(1-p)^2}$	$\dfrac{p}{(1-p)^2}$

Note regarding parameters: $\theta > 0\,,\ 0 < p < 1\,,\ q = 1-p\,,\ a = -\left(\ln(1-p)\right)^{-1}$

The probability and moment generating functions of the size-biased distributions arise directly from the corresponding generating functions of the original distributions. Generally, the following holds

$$G_{X^*}(t) = \sum_{x=1}^{k} t^x \pi^*_{x|\Theta} = \frac{1}{\mathrm{E}(X)} \sum_{x=1}^{k} x t^x \pi_{x|\Theta} = \frac{tG'_X(t)}{\mathrm{E}(X)} = \frac{tG'_X(t)}{G'_X(1)}, \tag{2.42}$$

$$M_{X^*}(t) = \sum_{x=1}^{k} e^{tx} \pi^*_{x|\Theta} = \frac{1}{\mathrm{E}(X)} \sum_{x=1}^{k} x e^{tx} \pi_{x|\Theta} = \frac{M'_X(t)}{\mathrm{E}(X)} = \frac{M'_X(t)}{M'_X(0)}, \tag{2.43}$$

wherefrom the cumulant generating function results as

$$K_{X^*}(t) = \ln M_{X^*}(t) = \ln M'_X(t) - \ln M'_X(0)\,. \tag{2.44}$$

Table 2.5 provides probability and moment generating functions of the discrete distributions listed in Table 2.4 for both original and its size-biased version.

Table 2.5: Probability and moment generating functions of some basic discrete distributions and their size-biased forms

Distribution	X		X^*	
	$G_X(t)$	$M_X(t)$	$G_{X^*}(t)$	$M_{X^*}(t)$
Poisson	$e^{\theta(t-1)}$	$e^{\theta(e^t-1)}$	$te^{\theta(t-1)}$	$e^{\theta(e^t-1)+t}$
Binomial	$(pt+q)^n$	$(pe^t+q)^n$	$t(pt+q)^{n-1}$	$e^t(pe^t+q)^{n-1}$
Negative Binomial	$p^k/(1-qt)^k$	$p^k/(1-qe^t)^k$	$tp^{k+1}/(1-qt)^{k+1}$	$e^t p^k/(1-qe^t)^{k+1}$
Logarithmic	$-a\ln(1-pt)$	$-a\ln(1-pe^t)$	$(1-p)t/(1-pt)$	$(1-p)e^t/(1-pe^t)$

Logarithmic distribution: $a = -\left(\ln(1-p)\right)^{-1}$

The k-th factorial moment of the size-biased distribution can be expressed by appropriate factorial moments of the original distribution. By substituting $t = 1$ in the k-th derivative of $G_{X^*}(t)$

$$G^{(k)}_{X^*}(t) = \frac{1}{\mu}\left(kG^{(k)}_X(t) + tG^{(k+1)}_X(t)\right)\,,$$

we obtain the following relation

$$\mu^*_{(k)} = k\frac{\mu_{(k)}}{\mu} + \frac{\mu_{(k+1)}}{\mu}\,, \quad k = 1, 2, \ldots \tag{2.45}$$

Also, the k-th raw moment of the size-biased random variable X^* is simply calculated as a ratio of the $(k+1)$-th raw moment and the mean of the original distribution

$$\mu^{*\prime}_k = \mathrm{E}((X^*)^k) = \sum_{x=1}^{k} x^k \pi^*_{x|\Theta} = \frac{1}{\mathrm{E}(X)} \sum_{x=1}^{k} x^{k+1} \pi_{x|\Theta} = \frac{\mathrm{E}(X^{k+1})}{\mathrm{E}(X)} = \frac{\mu'_{k+1}}{\mu}\,. \tag{2.46}$$

Consequently, we have $\mathrm{E}(X^*) = \mathrm{var}(X)/\mathrm{E}(X) + \mathrm{E}(X)$ for non-degenerate X, hence the size-biased mean is displaced for the value $\mathrm{var}(X)/\mathrm{E}(X)$[6] with respect to the original mean. Finally, central moments of the size-biased distribution can be derived directly from its raw moments by applying equations (B.13), Appendix B. Thus the variance of X^* becomes

$$\mathrm{var}(X^*) = \mu_2^{*\prime} - (\mu^*)^2 = \frac{\mu_3'\mu - (\mu_2')^2}{\mu^2}\,. \tag{2.47}$$

It is important to mention that the size-biased family (2.41) contains distributions with index of dispersion greater than, equal to and smaller than one. This feature enables them to fit discrete data in under-, equi- and overdispersed situations, as demonstrated in forthcoming chapters.

The more general case, known as *size-biased distributions of order* α, which corresponds to the weight $w(x, \alpha) = x^\alpha$ (cf. Rao, 1965) has also been widely utilized in statistics. Clearly, distribution (2.41) is its special case that results for $\alpha = 1$. Area-biased sampling applied extensively in forestry and discussed by Gove (2003a) is obtained for $\alpha = 2$. Individuals (standing or dead and fallen material in the forest) are selected into the sample with probability proportional to some arial attribute, the well-known example is tree basal area.

2.4 Parameter Estimation Procedures

Let $X_1, \ldots, X_n$ be a sample from a population having pmf $\pi_{x|\Theta} = P_\Theta(X = x)$ with an unknown parameter vector $\mathbf{\Theta} = (\theta_1, \ldots, \theta_q)$ taking on values in the parameter space Ω. The knowledge of parameter vector $\mathbf{\Theta}$ yields the knowledge of the entire population, thus our interest is to find a good estimator of $\mathbf{\Theta}$. An *estimator* is a function of the random variables $X_1, \ldots, X_n$ of a sample and should be distinguished from an *estimate*, a function of the realized values $x_1, \ldots, x_n$, which is a number. Sometimes we have to deal with quite complicated models requiring not always the same estimation procedures. Application of different estimation techniques yields very often different estimators. In this section we introduce three most common estimation methods: method of moments, maximum likelihood method and estimation based on sample mean and first frequency class.

2.4.1 Estimation by Method of Moments

The method of moments is one of the simplest estimation procedures that almost always yields some kind of estimates, provided the theoretical moments exist. It is based on equating the first q sample raw moments m_r', $r = 1, \ldots, q$, to the corresponding q population moments μ_r', where q is the number of unknown parameters. The population moments μ_r' are functions of unknown parameters $\theta_1, \ldots, \theta_q$ which

[6] Note that $\mathrm{var}(X)/\mathrm{E}(X)$ is the index of dispersion regarding the standard Poisson distribution.

are to be replaced by their empirical counterparts. Hence, the moment (MM) estimator $(\hat{\theta}_1, \ldots, \hat{\theta}_q)$ of $(\theta_1, \ldots, \theta_q)$ is obtained by solving simultaneously the following system of q equations

$$\begin{aligned} m'_1 &= \mu'_1(\hat{\theta}_1, \ldots, \hat{\theta}_q)\,, \\ &\vdots \\ m'_q &= \mu'_q(\hat{\theta}_1, \ldots, \hat{\theta}_q). \end{aligned} \tag{2.48}$$

Sometimes it is preferred to equate the first q central moments m_r, to the corresponding q population moments μ_r, or even the first q factorial moments $m_{(r)}$ can be set equal to the corresponding q population moments $\mu_{(r)}$. All three procedures yield identical estimates. Thereby the sample moments are calculated as follows

$$\begin{aligned} m'_r &= \frac{1}{n}\sum_{i=1}^{k} i^r f_i\,, \quad m_r = \frac{1}{n-1}\sum_{i=1}^{k}(i - m'_1)^r f_i\,, \\ m_{(r)} &= \frac{1}{n}\sum_{i=1}^{k} i(i-1)(i-2)\ldots(i-r+1)f_i\,, \end{aligned} \tag{2.49}$$

where n is the sample size, f_i denotes the absolute frequency of the i-th class, whereas the last frequency class is labelled by k. Also, note that $m'_1 = \bar{x}$ and $m_2 = s^2$.

Nevertheless, the range of the moment estimator may sometimes differ from the range of the parameter it is estimating. For example, we can get negative moment estimates of both binomial parameters[7], although per definition they should be positive. This can happen if and only if the sampled data exhibit overdispersion, i.e. the sample mean is smaller than the sample variance (cf. Neubauer and Friedl, 2006). Although the method of moments yields unbiased and consistent point estimators for moments, it is rather inefficient for functions of moments, which are generally biased and may have large quadratic errors. Therefore, we use moment estimates sometimes as starting values from which more efficient estimates can be determined. Basic properties of the estimators and the criteria for choosing the most appropriate one are given in Appendix C.

The *size-biased moment equations* require size-biased moments. For example, we can equate the raw moments given in equation (2.46) to the corresponding sample moments, obtaining as a solution the size-biased moment estimators.

2.4.2 Estimation by Maximum Likelihood

Consider a random sample of size n from a population with pmf $\pi_{i|\Theta} = P_\Theta(X = i)$. Let f_i denote the observed frequency of different classes i, $i = 1, \ldots, k$ such that $\sum_{i=1}^{k} f_i = n$, where k is the largest frequency class. The maximum likelihood (ML)

[7] The MM estimates of the binomial parameters are $\hat{n}_{\mathrm{MM}} = \bar{x}^2/(\bar{x} - s^2)$ and $\hat{p}_{\mathrm{MM}} = 1 - s^2/\bar{x}$, where $\bar{x}$ and s^2 denote the first two sample moments.

estimator of parameter vector $\mathbf{\Theta} = (\theta_1, \theta_2, \ldots, \theta_q)$ is the value that maximizes the likelihood function $L(\mathbf{\Theta}|f_1, \ldots, f_k)$ defined by

$$L(\mathbf{\Theta}|f_1, \ldots, f_k) = \prod_{i=1}^{k} \left(P_\Theta(X = i)\right)^{f_i} . \tag{2.50}$$

However, it is rather its log-likelihood function given by

$$l(\mathbf{\Theta}|f_1, \ldots, f_k) = \log L(\mathbf{\Theta}|f_1, \ldots, f_k) = \sum_{i=1}^{k} f_i \log P_\Theta(X = i) \tag{2.51}$$

that will be maximized, since maximizing the logarithm often requires simpler computation. Also, the logarithm is a strictly increasing function over the range of the likelihood, hence the two maxima coincide. Assuming $l(\mathbf{\Theta}|f_1, \ldots, f_k)$ to be differentiable in θ_i, the ML estimate $\widehat{\mathbf{\Theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)$ results as a solution of the score equations[8] $S(\mathbf{\Theta}) = 0$, where the score function is the following first derivative vector

$$S(\mathbf{\Theta}) = \frac{\partial l(\mathbf{\Theta}|f_1, \ldots, f_k)}{\partial \mathbf{\Theta}} = \left(\frac{\partial l(\mathbf{\Theta}|f_1, \ldots, f_k)}{\partial \theta_1}, \ldots, \frac{\partial l(\mathbf{\Theta}|f_1, \ldots, f_k)}{\partial \theta_q} \right)^{\mathrm{T}} . \tag{2.52}$$

Equations (2.52) can be solved iteratively for $\widehat{\mathbf{\Theta}}$ using the Newton-Raphson method (cf. Schwarz and Köckler, 2006) where the sequence $\{\mathbf{\Theta}^{(r)}, r = 1, 2, \ldots\}$ is determined from the recurrence relation

$$\mathbf{\Theta}^{(r+1)} = \mathbf{\Theta}^{(r)} - \frac{S(\mathbf{\Theta}^{(r)})}{H(\mathbf{\Theta}^{(r)})}, \quad r = 0, 1, \ldots \tag{2.53}$$

Thereby, H is the Hessian, the squared matrix of the second-order partial derivatives of $l(\mathbf{\Theta}|f_1, \ldots, f_k)$ that is evaluated at $\mathbf{\Theta}^{(r)}$. Commonly, the moment estimator $\widehat{\mathbf{\Theta}}_{\text{MM}}$ is taken as the starting value $\mathbf{\Theta}^{(0)}$ for the iteration process. The iteration process ends at the m-th step, if the inequality $|\mathbf{\Theta}^{(m+1)} - \mathbf{\Theta}^{(m)}| < \varepsilon$ holds for some positive integer m and sufficiently small value of ε. Then, $\mathbf{\Theta}^{(m)}$ is the solution of the equation (2.52) and hence $\mathbf{\Theta}^{(m)} = \widehat{\mathbf{\Theta}}_{\text{ML}}$. The ML estimators are not always unbiased, however, if an efficient estimator exists, it will be achieved by the ML method, at least asymptotically. It means that as the sample size increases to infinity, the ML estimator achieves the Cramér-Rao lower bound (CRLB). For the definition see Appendix C. Also, ML estimators are consistent and asymptotically normal.

Furthermore, the variance-covariance matrix of $\widehat{\mathbf{\Theta}}_{\text{ML}}$ can be obtained by inverting the Fisher information matrix $\mathcal{I}(\widehat{\mathbf{\Theta}})$ defined as minus the Hessian of $l(\mathbf{\Theta}|f_1, \ldots, f_k)$ evaluated at $\widehat{\mathbf{\Theta}}_{\text{ML}}$. At the maximum the second derivative of $l(\mathbf{\Theta}|f_1, \ldots, f_k)$ is negative, thus $\mathcal{I}(\widehat{\mathbf{\Theta}})$ measures in some way the "sharpness" of the log-likelihood curve in

[8] Note that the solutions of the score equation are the only possible ML candidates, since the zeros of the first derivative yield extremes (local or global) or inflection points. Our interest is rather to find a global maximum. Thus, the matrix of second derivatives has to be negative definite for all values of $\mathbf{\Theta}$.

the neighbourhood of $\widehat{\mathbf{\Theta}}_{\mathrm{ML}}$. Low information means flat maximum, hence indicates more uncertainty about $\mathbf{\Theta}$ and vice versa. Parameters with larger information will have smaller variance, therefore greater precision. An approximate 95% confidence interval for the parameter θ_i is obtained as

$$\hat{\theta}_i \pm 1.96\sqrt{\mathrm{var}(\hat{\theta}_i)}\,, \tag{2.54}$$

where $\mathrm{var}(\hat{\theta}_i)$ is the i-th diagonal element of the squared matrix $(\mathcal{I}(\widehat{\mathbf{\Theta}}))^{-1}$.
In some situations the *invariance property* of the ML estimate can be useful. It says that if $\widehat{\mathbf{\Theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_q)$ is the ML estimate of $\mathbf{\Theta} = (\theta_1, \dots, \theta_q)$ and $g(\theta_1, \dots, \theta_q)$ is any function of the parameters, then the ML estimate of $g(\theta_1, \dots, \theta_q)$ is $g(\hat{\theta}_1, \dots, \hat{\theta}_q)$. For further references see e.g. Casella and Berger (2002) or Pawitan (2001).

The general form of the *size-biased likelihood*, independent of any distribution assumption, is given by

$$L^*(\mathbf{\Theta}|f_1, \dots, f_k) = \prod_{i=1}^{k} (P_\Theta(X^* = i))^{f_i} = \prod_{i=1}^{k} \left(\frac{iP_\Theta(X = i)}{\mathrm{E}(X)}\right)^{f_i}, \tag{2.55}$$

hence the log-likelihood function results in

$$l^*(\mathbf{\Theta}|f_1, \dots, f_k) = \sum_{i=1}^{k} \Big(f_i \log i + f_i \log P_\Theta(X = i) - f_i \log \mathrm{E}(X)\Big)\,. \tag{2.56}$$

Notice that the first term in the formula above is a constant which depends only on the data and thus may be dropped if needed, the second term is the log-likelihood of the original distribution given in (2.51) and the last term $\sum_{i=1}^{k} f_i \log \mathrm{E}(X) = n \log \mu$ is a correction term, emphasizing the fact that the observations were not sampled with equal probability (cf. Gove, 2003a, 2003b).

In order to apply the Newton-Raphson iteration algorithm the score function of the first derivatives of $l^*(\mathbf{\Theta}|f_1, \dots, f_k)$ is required. Hence, the score equations are

$$\frac{\partial l^*(\mathbf{\Theta}|f_1, \dots, f_k)}{\partial \theta_i} = \frac{\partial l(\mathbf{\Theta}|f_1, \dots, f_k)}{\partial \theta_i} - \frac{n}{\mu}\frac{\partial \mu}{\partial \theta_i} = 0\,. \tag{2.57}$$

These will be combined with the second-order partial derivatives given by

$$\begin{aligned}
\frac{\partial^2 l^*(\mathbf{\Theta}|f_1, \dots, f_k)}{\partial \theta_i^2} &= \frac{\partial^2 l(\mathbf{\Theta}|f_1, \dots, f_k)}{\partial \theta_i^2} + \frac{n}{\mu^2}\left(\frac{\partial \mu}{\partial \theta_i}\right)^2 - \frac{n}{\mu}\frac{\partial^2 \mu}{\partial \theta_i^2}\,,\\
\frac{\partial^2 l^*(\mathbf{\Theta}|f_1, \dots, f_k)}{\partial \theta_i \partial \theta_j} &= \frac{\partial^2 l(\mathbf{\Theta}|f_1, \dots, f_k)}{\partial \theta_i \partial \theta_j} + \frac{n}{\mu^2}\frac{\partial \mu}{\partial \theta_j}\frac{\partial \mu}{\partial \theta_i} - \frac{n}{\mu}\frac{\partial^2 \mu}{\partial \theta_i \partial \theta_j}\,.
\end{aligned} \tag{2.58}$$

The score vector and the Hessian matrix are of the same form as the log-likelihood, being composed of the part from original distribution and correction components. Interestingly enough, the correction term depends on the size-biased order α because in the general case $\mu = \sum_{x=1}^{k} x^{\alpha}\pi_{x|\Theta}$. Hence, unique corrections with respect to size-biased (α=1) and area-biased (α=2) log-likelihoods, score and Hessian exist.

In practice, we often use numerical optimization procedures to find $\widehat{\mathbf{\Theta}}$ directly from the log-likelihood function. In that case we do not need to find analytically the score function or the Hessian, since they are provided numerically by the procedure. The variance-covariance matrix, as well as confidence intervals of the estimated parameters, can afterwards be easily calculated from the resulting Hessian, as explained above.

2.4.3 Estimation Based on Mean and First Frequency Class

This method has proven to be useful for various situations in which the frequency of the lowest class in the sample is much larger than the other frequency classes or when the graph of the sample distribution is approximately L-shaped (cf. Anscombe, 1950; Johnson and Kotz, 1969). It has been already mentioned in Section 1.8 that the lowest frequency class is the first one, since zero-syllable words as parts of the subsequent words do not exist as a separate word class. Therefore, the approach here is to equate the sample mean $\bar{x}$ and the relative frequency of the first class f_1/n to the population mean μ and the probability of the first class $\pi^{d}_{1|\Theta}$ (or $\pi^{*}_{1|\Theta}$ in size-biased case), respectively, in order to give more weight to this large frequency class. The estimators are obtained by solving the resulting system of simultaneous equations and are denoted by FF. This technique offers an efficient and useful alternative to the maximum likelihood method. Moreover, it is likely to give better results, with increasing values of the corresponding $\pi_{x|\Theta}$'s.

2.5 Sampling From Discrete Probability Models

In this section we give a short introduction to the inversion algorithm, discussed by Stadlober (1989, 7ff.), needed for generating random variables of the distributions considered in Chapters 4 - 7. The inversion concept is based on the following theorem.

Theorem 2.2 *Let $F(x)$ be any distribution function on $\mathbb{R}$. If U is uniformly distributed on $(0,1)$, then the random variable $X = inf\{y|U \leq F(y)\}$ has distribution function $F(x)$.*

Next theorem gives a specialization to discrete random variables.

Corollary 2.1 *Let $F(x)$ be a discrete cdf of a distribution $\{\pi_j\}$, $j = 0, 1, \ldots$ Define*

$$X = k\,, \quad if \quad \sum_{i=0}^{k-1} p_i < U \leq \sum_{i=0}^{k} p_i\,, \quad k = 0, 1, \ldots\,, \tag{2.59}$$

where $U \sim U(0,1)$. Then X has distribution function $F(x)$.

Assume that there is a recurrence relation for the probabilities of the discrete distribution of interest defined by

$$\pi_k = g(\pi_{k-1})\,, \quad k = 2, 3, \ldots \tag{2.60}$$

The discrete random variables are generated by using the following search procedure:

0. [Pre-set] Define probability π_1 and precision b.
1. Generate $U \sim U(0,1)$. Set $K \leftarrow 1$, $f \leftarrow \pi_1$.
2. If $U \leq f$ return K.
3. If $K > b$ go to step 1.
4. Set $U \leftarrow U - f$, $K \leftarrow K + 1$, $f \leftarrow g(f)$ and go to step 2.

In step 3. we do not allow K to exceed the bound b at which we are sure that the calculated probabilities π_k are insignificant. Hence, b is found as $b = \min\{k|\pi_k < \varepsilon\}$, with $\varepsilon \in \{10^{-4}, 10^{-5}, 10^{-6}, \ldots\}$. Obviously, different choices of ε yield also different bounds of b. This method is also known as a *chop-down search* and is applicable when the distribution has a small number of mass points.

2.6 Goodness of Fit

Whenever count data are considered and one wants to test the fit of a certain model, the standard procedure is to compare the expected frequencies with the observed frequencies. It means testing the null hypothesis

$$H_0 : \pi_i = \pi_{0i}(\mathbf{\Theta})\,, \quad i = 1, 2, \ldots, k\,, \quad \mathbf{\Theta} \in \Omega \subset \mathbb{R}^q \tag{2.61}$$

that the observed frequency distribution is consistent with a particular theoretical distribution. Here, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ denotes the true probability vector of the event that some particular word has length i, whereas $\boldsymbol{\pi}_\mathbf{0} = (\pi_{01}, \pi_{02}, \ldots \pi_{0k})$ is the hypothesized probability vector evaluated at $\mathbf{\Theta} = (\theta_1, \ldots, \theta_q)$ with $q < k - 1$. The probability of the largest word length class π_k, is calculated as 1 minus the sum of all remaining classes, since there are no infinitely long words in the language. To verify the hypothesis that a certain model holds we calculate Pearson's X^2 test statistics defined by

$$X^2 = \sum_{i=i}^{k} \frac{(f_i - n\pi_{0i}(\mathbf{\Theta}))^2}{n\pi_{0i}(\mathbf{\Theta})}\,, \quad i = 1, 2, \ldots, k\,, \tag{2.62}$$

where the vector $\mathbf{f} = (f_1, \ldots, f_k)$ is the number of words of length i, and n represents the total number of words in a given text. All unspecified parameters θ_j, $j = 1, \ldots, q$ have to be estimated before the expected frequencies $n\pi_{0i}(\mathbf{\Theta})$ are calculated. If the observed vector $\mathbf{f} = (f_1, \ldots, f_k)$ is replaced by the random vector $\mathbf{F} = (F_1, \ldots, F_k)$, then X^2 is a random variable which is $X^2 \overset{\text{as}}{\sim} \chi^2_{k-q-1}$ distributed when H_0 is true. Therefore, the null hypothesis will be rejected at the significance level α whenever $X^2 \geq \chi^2_{k-q-1;1-\alpha}$ holds.

Furthermore, Pearson's X^2 test statistics is a member of the *power-divergence family*[9] of the goodness-of-fit statistics, introduced by Cressie and Read (1984, 1988). It examines, like other test statistics of this family, the adequacy of the model by evaluating the discrepancy between observed frequencies and their hypothetical expectations, however it is strongly dependent on the sample size. If the difference between the theoretical and empirical frequencies is taken to be fixed, all these statistics increase linearly with the increase of the sample size (cf. Grotjahn and Altmann, 1993). It means that in case of large sample sizes, even small model deviations will be detected as significant having as a consequence the possible model rejection. Another disadvantage of the goodness-of-fit test is that it indicates only a statistical significance of a possible deviation from the model, but not the order of the divergence. Therefore, a direct comparison of goodness-of-fit values in case of different sample sizes may be misleading. Also, differences in the number of classes influence differences in the degrees of freedom.

2.6.1 Discrepancy Index

In order to avoid the problems with the goodness-of-fit test it was necessary to find some measure allowing to compare fits of a single model to empirical samples differing in the sample size. To measure the degree of the goodness of fit, Moore (1984) considered the value T/n, where T denotes any of the statistics belonging to the power-divergence family. The usage of the standardized discrepancy index $C = X^2/n$, as its special case, became common practice in linguistics (cf. Grotjahn and Altmann, 1993). Obviously, the relation (2.62) can be written as

$$X^2 = \sum_{i=1}^{k} \frac{\left(F_i - n\pi_{0i}(\widehat{\boldsymbol{\Theta}})\right)^2}{n\pi_{0i}(\widehat{\boldsymbol{\Theta}})} = n \sum_{i=1}^{k} \frac{\left(F_i/n - \pi_{0i}(\widehat{\boldsymbol{\Theta}})\right)^2}{\pi_{0i}(\widehat{\boldsymbol{\Theta}})},$$

when the observed vector $\boldsymbol{f}$ is replaced by the random vector $\boldsymbol{F}$. Evidentially, the value $C = X^2/n$, called the *standardized discrepancy index*, measures discrepancy between the empirical relative frequencies F/n and the probabilities $\boldsymbol{\pi}_0(\widehat{\boldsymbol{\Theta}})$ estimated under H_0. The discrepancy index has the advantage that it does not depend on the degrees of freedom and sample size. At the beginning of empirical research, the values of $C > 0.02$ were usually considered unacceptable and hence the model rejection was recommended (cf. Antić et al., 2005, 2006b). Mačutek, Švehlíková, and Cenkerová (2011) studied the rank-frequency distributions of melodic intervals and showed that fitting the negative hypergeometric distribution to the data yields

[9] This family includes also the log-likelihood ratio statistics, the Freeman-Tukey statistics, the modified log-likelihood ratio statistics and the Neyman-modified statistics as its special cases.

contradictory results. Although the fit proved not to be satisfactory in terms of discrepancy index C, it was considered very good in terms of determination coefficient R^2. Since rejection rules for both C and R^2 are not theoretically derived, but only considered as rules of thumb, the above authors suggested an extensive future study, based on many data sets from different areas, in order to get at least roughly comparable results. Based on empirical investigations done in the framework of QuanTA[10] project, we consider here the fit of the model (a) extremely good if $C \leq 0.01$, (b) good if $0.01 < C \leq 0.02$ and (c) acceptable if $0.02 < C \leq 0.05$ (cf. Djuraš and Stadlober, 2010). However, there is a need for an extensive simulation study to evaluate the plausibility of the suggestions above.

2.7 Software Used

All computations regarding parameter estimation and verifications of the goodness of fits done in this work were mainly implemented by the public domain statistical software `R`, `version 2.8.0` (cf. R Development Core Team, 2008). The only exception are the distributions arising as special cases of the Fucks' generalized Poisson distribution, dealt with in Chapter 3. For the corresponding calculations in this area we used the software `Maple`, `version 9` (cf. Ablamowicz and Fauser, 2011). The main results concerning this part of calculations are summarized in Tables 3.6 - 3.9. Additionally, random variables from the distributions of interest were generated in R via the inversion method explained in Section 2.5 and subsequently simulation studies were done. Several R functions developed here are documented in Appendix D. Although the software R is an open source project, there are many books written already on this topic. A great deal of useful information can be obtained from the home page `http://www.r-project.org/`.

[10] More details considering the main results of this project were given in Chapter 1.

Chapter 3

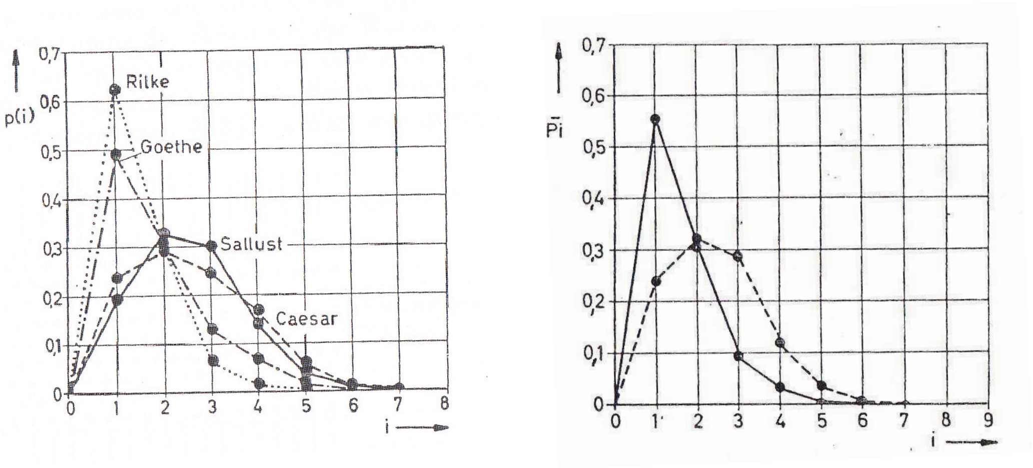
Fucks' Generalized Poisson Distribution

3.1 Historical Context

Scientific work of German physicist Wilhelm Fucks (1902–1990) motivated by the desire to find the "mathematical law of the process of word formation from syllables for all those languages, which form their words from syllables" (cf. Fucks, 1955) was remarkable in the history of word length studies and remains worth mentioning still today.

Based on elaborate mathematical ideas and the conditions that each word contains at least one syllable, Fucks (1955, p. 209) assumed that the 1-displaced Poisson distribution might be accepted as a general theoretical model for word length frequency distributions. Even more, he proved that his suggested model, often termed as the *Fucks' model* by the linguist community, is a special case of a much more general model. However, this generalization, named *Fucks' Generalized Poisson Distribution* (Fucks' GPD) throughout this chapter, has hardly ever been discussed in detail. Rather, he mentioned it only in a very few of his publications (cf. Fucks, 1956a, 1956c) but without application to the linguistic material. Interestingly, Fucks' GPD has been discussed more intensively by several scholars from Eastern Europe and the former Soviet Union, due to the Russian translation of Fucks' (1956c) article. In that context Fucks' theoretical assumptions were not only generally accepted and applied to specific linguistic material, but also served as a starting point for new developments with regard to alternative ways of parameter estimation and even further generalizations (cf. Section 3.3).

As to the study of text and language, Fucks' work includes the research on the individual styles of single authors, as well as the properties of texts from different authors, both from one single language but also across languages. As an example for the comparison of two different languages, distributions of relative frequencies of texts of two German authors, Rilke and Goethe, and of two Latin authors, Sallust and Caesar, given by Fucks (1957, p. 33), are displayed in Figure 3.1(a)[1]. Being convinced only by the graphical impressions, Fucks came to the conclusion that the curves of different authors but of one specific language show certain similarity. In contrary to this, curves of the German and Latin texts turned out to be quite different. The small differences within texts could be explained, according to him, as being characteristic for the author's individual style, rather than text specific. To eliminate this author specific differences for the texts written in one single language, Fucks suggested to look at the frequency distributions of a huge number of texts of various authors and languages by calculating the mean values



(a) German and Latin authors differences

(b) Language differences

Figure 3.1: Distribution of relative frequencies of syllables per word

$$\bar{p_i}^{(\lambda)} = \frac{1}{T}\sum_{t=1}^{T} p_i^{(t;\lambda)}\,, \quad 1 \leq i \leq I\,, \quad 1 \leq t \leq T\,, \tag{3.1}$$

where $p_i^{(t;\lambda)}$ denotes the relative frequency of the i-syllable word of the text t written in language λ. The analyzed languages are marked by $1 \leq \lambda \leq \Lambda$. Furthermore, the number of texts T should be considered to be large enough, in order to have a representative cross-section of the given language (cf. Fucks, 1955, p. 202). Figure 3.1(b), taken from Fucks (1957, p. 33), represents the results for a great number of German and Latin texts. Two curves, regarding to "the German" and "the Latin" author, are obviously of a different type.

Unfortunately, Fucks never presented any raw data, a fact which renders it impossible to control the results he obtained. Moreover, he did not apply any tests to

[1] Figure 3.1(a) shows relative frequency distributions of the following texts: Rilke *"Die Weise von Liebe und Tod des Cornets"*, Goethe *"Wilhelm Meisters Lehr- und Wanderjahre"*, Sallust *"Belleum Jugurthinum"* and Caesar *"De Bello Gallico"*.

check the statistical significance of the goodness of his fits. Being probably aware of the fact that the value of the X^2 test statistics, introduced in Section 2.6, linearly increases with an increase of the sample size and hence results in significant differences for larger samples, Fucks emphasized that his linguistic data are not particularly adequate for the application of the χ^2 test. Thus, one was forced to believe that the 1-displaced Poisson distribution is the best possible model on the basis of his empirical investigations. Only in the case when he compared texts from eight natural and one artificial languages, at least the relative frequencies are published (cf. Fucks, 1956a, 1956c). Original data from those nine different languages are presented in Table 3.1. In addition to the relative frequencies of i-syllable words calculated by formula (3.1), Table 3.1 contains also the mean word length $\bar{x}$, as well as the sample variance s^2, for each of the given languages. The latter two are corrected values from Fucks' original ones which are obtained from relative frequencies $\bar{p}_i$. Figure 3.2 illustrates word length frequency distributions of Fucks' data. Apparently, the curve for Arabic shows certain systematic deviation compared to eight other languages of a similar pattern. This could be explained by the fact that the observed data were gained from only two texts, both from the same genre, namely,

Table 3.1: Relative frequencies $\bar{p}_i$, mean and variance of nine languages (Fucks, 1956a)

i	English	German	Esperanto	Arabic	Greek	Japanese	Russian	Latin	Turkish
1	0.7152	0.5560	0.4040	0.2270	0.3760	0.3620	0.3390	0.2420	0.1880
2	0.1940	0.3080	0.3610	0.4970	0.3210	0.3440	0.3030	0.3210	0.3784
3	0.0680	0.0938	0.1770	0.2239	0.1680	0.1780	0.2140	0.2870	0.2704
4	0.0160	0.0335	0.0476	0.0506	0.0889	0.0868	0.0975	0.1168	0.1208
5	0.0056	0.0071	0.0082	0.0017	0.0346	0.0232	0.0358	0.0282	0.0360
6	0.0012	0.0014	0.0011	–	0.0083	0.0124	0.0101	0.0055	0.0056
7	–	0.0002	–	–	0.0007	0.0040	0.0015	0.0007	0.0004
8	–	0.0001	–	–	–	0.0004	0.0003	0.0002	0.0004
9						0.0004	–	–	–
$\bar{x}$	1.4064	1.6333	1.8971	2.1032	2.1106	2.1325	2.2268	2.3894	2.4588
s^2	0.5645	0.7442	0.8532	0.6579	1.3526	1.3952	1.4220	1.2093	1.1692

fairy tales (cf. Fucks, 1956a, p. 15), which can not be understood to be statistically representative random cross-section of Arabic texts. Although Fucks claimed that the 1-displaced Poisson model can be accepted as an overall valid standard model for word length frequencies, we will show in Section 3.5 that it fits only three of the nine observed languages above. One reason for this might be the fact that the data for each of the languages originated from text mixtures, not from individual texts. Therefore, the material might be characterized by an internal heterogeneity, violating the statistical principles of data homogeneity. However, this point can not be pursued in detail, hence it will be neglected here, and Fucks' data will be used as an exemplary material throughout the following discussions.

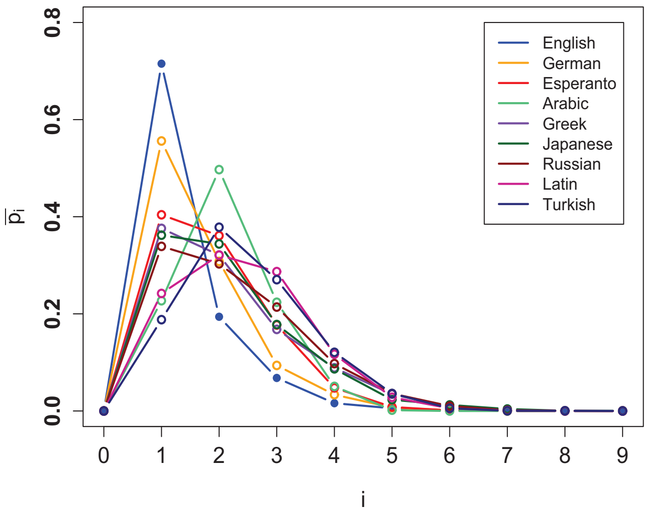


Figure 3.2: Distribution of relative frequencies of word lengths of Fucks' (1956a) data

The purpose of this chapter is to consider the concept of Fucks' GPD more precisely, as well as the modifications and generalizations derived from it. Also, different parameter estimation methods are discussed. Since the relevant works are not systematic in their approaches, the corresponding derivations will be calculated and presented here, in detail.

3.2 Derivation of the Fucks GPD

Trying to find a universal mechanism to describe the process of word formation out of syllables Fucks (1956a) assumed that this process can be seen as a superposition of two processes, one being completely random and the other completely determined. The final formula describing the probabilistic organization of this mathematical law is derived in the following way (see also the preliminary study of Antić et al., 2005).

Let us consider a large number of elements which should be stochastically distributed on a great number of cells. We distinguish between elements of two kinds, namely one- and zero-elements. It is further assumed that each of these cells has a place for n and only n elements. The probability for a one-element to fall into a specific cell at a given place within that cell will be denoted by q. The complementary probability for a zero-element will be thus $1 - q$. Before starting the distribution process a certain number of elements will be placed into a certain number of cells so that there will be $\varepsilon_k - \varepsilon_{k+1}$ cells pre-occupied with k one-elements in them, where the trivial condition $\varepsilon_k \geq \varepsilon_{k+1}$ holds. Once this pre-placing is completed, we randomly distribute the rest of the elements on the remaining $n - k$ empty places in the cells. Hence, the probability to find further $i - k$ one-elements and $n - i$ zero-elements in one given cell, in some specific order is $q^{i-k}(1 - q)^{n-i}$ while the probability of all possible arrangements is $\binom{n-k}{i-k}$.

As a result of these considerations Fucks (1956a, p. 12) finally conceived a model, giving the probability to find exactly i one-elements in a given cell. This model later became known as the *Fucks binomial distribution* and is defined as

$$\pi^{\text{FB}}_{i|q,\varepsilon_k} = P(X = i) = \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \begin{pmatrix} n-k \\ i-k \end{pmatrix} q^{i-k}(1-q)^{n-i}\,, \quad i = 0, \ldots, n. \tag{3.2}$$

The random variable X denotes the number of syllables per word, $\pi^{\text{FB}}_{i|q,\varepsilon_k}$ is the probability that a given word has i syllables, with $\sum_i \pi^{\text{FB}}_{i|q,\varepsilon_k} = 1$, $q = 1-p$, $0 < p < 1$, $n \in \mathbb{N}$. Ultimately, the distribution (3.2) is a sum of weighted binomial probabilities. The corresponding weights are labelled as $(\varepsilon_k - \varepsilon_{k+1})$, k indicating the number of components to be analyzed. The expected number of syllables per word or mean of the distribution (3.2) is found to be

$$\mu = \sum_{i=1}^{n} i\pi^{\text{FB}}_{i|q,\varepsilon_k} = (n - \varepsilon')q + \varepsilon' \text{ with } \varepsilon' = \sum_{k=1}^{\infty} \varepsilon_k\,, \tag{3.3}$$

and can be obtained by decomposition of the summation into corresponding ε terms

$$\begin{aligned}
\sum_{i=1}^{n} i\pi^{\text{FB}}_{i|q,\varepsilon_k} &= \sum_{i=1}^{n} i \left(\sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \begin{pmatrix} n-k \\ i-k \end{pmatrix} q^{i-k}(1-q)^{n-i} \right) \\
&= \varepsilon_0 \sum_{i=0}^{n} iq^i(1-q)^{n-i} \binom{n}{i} + \\
&+ \varepsilon_1 \left(\sum_{i=1}^{n} iq^{i-1}(1-q)^{n-i} \binom{n-1}{i-1} - \sum_{i=1}^{n} iq^i(1-q)^{n-i} \binom{n}{i} \right) + \\
&+ \varepsilon_2 \left(\sum_{i=2}^{n} iq^{i-2}(1-q)^{n-i} \binom{n-2}{i-2} - \sum_{i=1}^{n} iq^{i-1}(1-q)^{n-i} \binom{n-1}{i-1} \right) + \ldots \\
&= qn + \varepsilon_1(1-q) + \varepsilon_2(1-q) + \ldots = qn + (1-q) \sum_{k=1}^{\infty} \varepsilon_k = qn + (1-q)\varepsilon'\,.
\end{aligned}$$

When $n \to \infty$ and $q \to 0$ such that $\mu - \varepsilon' = (n - \varepsilon')q = const$ the probability (3.2) can be written in the form

$$\begin{aligned} \pi^{\mathrm{FB}}_{i|q,\varepsilon_k} &= \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \binom{n-k}{i-k} \left(\frac{\mu - \varepsilon'}{n - \varepsilon'}\right)^{i-k} \left(1 - \frac{\mu - \varepsilon'}{n - \varepsilon'}\right)^{n-i} \\ &= \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \frac{(\mu - \varepsilon')^{i-k}}{(i-k)!} \frac{(n-k)!}{(n-i)!(n-\varepsilon')^{i-k}} \left(1 - \frac{\mu - \varepsilon'}{n - \varepsilon'}\right)^{n-i} . \end{aligned} \tag{3.4}$$

We use the fact that

$$\begin{aligned} \frac{(n-k)!}{(n-i)!(n-\varepsilon')^{i-k}} &= \frac{(n-k)(n-k-1)\dots(n-i+1)(n-i)!}{(n-i)!(n-\varepsilon')^{i-k}} \\ &= \left(1 - \frac{k - \varepsilon'}{n - \varepsilon'}\right) \left(1 - \frac{k - \varepsilon' + 1}{n - \varepsilon'}\right) \dots \left(1 - \frac{i - \varepsilon' - 1}{n - \varepsilon'}\right) \end{aligned} \tag{3.5}$$

tends to unity if n increases without limit and apply the equality

$$\left(1 - \frac{\mu - \varepsilon'}{n - \varepsilon'}\right)^{n-i} = \left(\left(1 + \frac{1}{\frac{n-\varepsilon'}{-(\mu-\varepsilon')}}\right)^{\frac{n-\varepsilon'}{-(\mu-\varepsilon')}} \right)^{-(\mu-\varepsilon')} \left(1 - \frac{\mu - \varepsilon'}{n - \varepsilon'}\right)^{\varepsilon'-i} \tag{3.6}$$

where

$$\left(\left(1 + \frac{1}{\frac{n-\varepsilon'}{-(\mu-\varepsilon')}}\right)^{\frac{n-\varepsilon'}{-(\mu-\varepsilon')}} \right)^{-(\mu-\varepsilon')} \to e^{-(\mu-\varepsilon')} \text{ as } n \to \infty$$

and

$$\left(1 - \frac{\mu - \varepsilon'}{n - \varepsilon'}\right)^{\varepsilon'-i} \to 1 \text{ as } n \to \infty .$$

Hence, if $n \to \infty$ and $q \to 0$ such that $\mu - \varepsilon' = (n - \varepsilon')q = const$ the Fucks binomial distribution (3.2) converges to the *Fucks' GPD*, which by replacing $(\mu - \varepsilon')$ with λ has the following form

$$\pi^{\mathrm{FGP}}_{i|\lambda,\varepsilon_k} = P(X = i) = e^{-\lambda} \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \frac{\lambda^{i-k}}{(i-k)!} , \quad i = 0, 1, 2, \dots \tag{3.7}$$

The condition $\lambda = \mu - \varepsilon' \geq 0$ implying $\mu \geq \varepsilon'$ is required to ensure nonnegative probabilities. In fact, the distribution (3.7) is the sum of weighted Poisson probabilities with parameter λ. The parameters $\{\varepsilon_k\}$ of the distribution are called the *ε-spectrum*. For probabilities in (3.7), the following conditions are additionally postulated by Fucks (1956a, p. 12):

(a) the condition $\varepsilon_k - \varepsilon_{k+1} \geq 0$ implies that $\varepsilon_k \geq \varepsilon_{k+1}$;

(b) since the sum of all the weights equals 1, we have $1 = \sum_{k=0}^{\infty}(\varepsilon_k - \varepsilon_{k+1}) = \sum_{k=0}^{\infty}\varepsilon_k - \sum_{k=0}^{\infty}\varepsilon_{k+1} = \varepsilon_0$;

(c) from (a) and (b), we obtain a more restrictive condition as to the ε-spectrum requiring $1 = \varepsilon_0 \geq \varepsilon_1 \geq \varepsilon_2 \geq \ldots \geq \varepsilon_k \geq \varepsilon_{k+1} \geq \ldots$

The following investigations use conditions (a) to (c) as basic restrictions. However, assuming further that some $\varepsilon_l = 0$ for $k \geq l$, then based on condition (a) it results in $\varepsilon_{l-1} \geq 0$, and hence $1 \geq \varepsilon_1 \geq \ldots \geq \varepsilon_{l-1} \geq 0$. In Section 3.4, where the special cases of the Fucks' GPD model are discussed, this additional restriction is required.

In the following section, we go more into details, concentrating on the derivation of the probability generating function and moments of the Fucks' GPD. Also, our aim is to determine the unknown ε_k values which characterize the Fucks' GPD as given in equation (3.7).

3.2.1 Probability Generating Function and Moments

The mean of the Fucks Generalized Poisson model (3.7) can be derived directly, by definition, which after summing up the corresponding ε expressions results in

$$\begin{aligned}
\mathrm{E}(X) = \sum_{i=1}^{\infty} i\pi_{i|\lambda,\varepsilon_k}^{\text{FGP}} &= e^{-\lambda}\sum_{i=1}^{\infty} i\left(\sum_{k=0}^{\infty}(\varepsilon_k - \varepsilon_{k+1})\frac{\lambda^{i-k}}{(i-k)!}\right) \\
&= e^{-\lambda}\left(\sum_{i=1}^{\infty} i\frac{\lambda^i}{i!} + \varepsilon_1\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} + \varepsilon_2\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} + \ldots\right) \qquad (3.8) \\
&= e^{-\lambda}\left(\lambda e^{\lambda} + e^{\lambda}\sum_{k=1}^{\infty}\varepsilon_k\right) = \lambda + \varepsilon' = \mu\,.
\end{aligned}$$

This fact is consistent with the border condition when the Fucks binomial distribution converges to the Fucks GPD. To determine $\mathrm{E}(X^2)$ we calculate

$$\begin{aligned}
\sum_{i=1}^{\infty} i^2\pi_{i|\lambda,\varepsilon_k}^{\text{FGP}} &= e^{-\lambda}\sum_{i=1}^{\infty} i^2\left(\sum_{k=0}^{\infty}(\varepsilon_k - \varepsilon_{k+1})\frac{\lambda^{i-k}}{(i-k)!}\right) \\
&= e^{-\lambda}\left(\sum_{i=1}^{\infty} i^2\frac{\lambda^i}{i!} + \varepsilon_1\sum_{i=0}^{\infty}(2i+1)\frac{\lambda^i}{i!} + \varepsilon_2\sum_{i=0}^{\infty}(2i+3)\frac{\lambda^i}{i!} + \ldots\right) \\
&= e^{-\lambda}\left(\sum_{i=1}^{\infty} i^2\frac{\lambda^i}{i!} + 2\sum_{i=1}^{\infty} i\frac{\lambda^i}{i!}\sum_{k=1}^{\infty}\varepsilon_k + \sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\sum_{k=1}^{\infty}(2k-1)\varepsilon_k\right).
\end{aligned}$$

Simplification of the expression above gives

$$\mathrm{E}(X^2) = \lambda(\lambda+1) + 2\varepsilon'\lambda + \sum_{k=1}^{\infty}(2k-1)\varepsilon_k\,,$$

which substituting $\lambda = \mu - \varepsilon'$ yields

$$\mathrm{E}(X^2) = \mu^2 + \mu - \varepsilon'^2 + 2\sum_{k=1}^{\infty}(k-1)\varepsilon_k\,.$$

Therefore, the variance of the Fucks GPD becomes

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X) = \mu - \varepsilon'^2 + 2\sum_{k=1}^{\infty}(k-1)\varepsilon_k\,. \tag{3.9}$$

The probability generating function (pgf) of the Fucks GPD random variable can be found directly by the use of definition (B.1), Appendix B, and the series expansion of $e^x = \sum_{i=0}^{\infty} x^i/i!$ for $x = \lambda t$. Hence, we have

$$G(t) = e^{\lambda(t-1)}\sum_{k=0}^{\infty}(\varepsilon_k - \varepsilon_{k+1})t^k\,. \tag{3.10}$$

Now, based on the pgf we can easily derive all factorial moments, denoted by $\mu_{(k)}$, of the distribution (3.7). However, for many linguistic problems it is sufficient to indicate only a few moments of the given distribution. For extensive explanation of the relation between the pgf and the moments of the distribution see Appendix B. Differentiating (3.10) partially with respect to t we have

$$\frac{\partial G(t)}{\partial t} = e^{\lambda(t-1)}\lambda\sum_{k=0}^{\infty}(\varepsilon_k - \varepsilon_{k+1})t^k + e^{\lambda(t-1)}\sum_{k=0}^{\infty}kt^{k-1}(\varepsilon_k - \varepsilon_{k+1})\,, \tag{3.11}$$

which by substituting $t = 1$ and using the relations

$$\sum_{k=0}^{\infty}(\varepsilon_k - \varepsilon_{k+1}) = 1 \;\text{ and }\; \sum_{k=0}^{\infty}k(\varepsilon_k - \varepsilon_{k+1}) = \sum_{k=1}^{\infty}\varepsilon_k = \varepsilon'\,, \tag{3.12}$$

gives the first factorial moment $\mu_{(1)}$, which is in fact the mean of the distribution

$$\mu_{(1)} = \lambda + \varepsilon' = \mu\,. \tag{3.13}$$

The second factorial moment $\mu_{(2)}$ arises from the second derivative of $G(t)$ with respect to t, putting $t = 1$ and using (3.12) together with the relation

$$\sum_{k=0}^{\infty}k^2(\varepsilon_k - \varepsilon_{k+1}) = \sum_{k=1}^{\infty}(2k-1)\varepsilon_k\,. \tag{3.14}$$

Hence,

$$\begin{aligned}\frac{\partial^2 G(t)}{\partial t^2} =& \lambda^2 e^{\lambda(t-1)}\sum_{k=0}^{\infty}(\varepsilon_k - \varepsilon_{k+1})t^k + 2\lambda e^{\lambda(t-1)}\sum_{k=0}^{\infty}k(\varepsilon_k - \varepsilon_{k+1})t^{k-1}\\ &+ e^{\lambda(t-1)}\sum_{k=0}^{\infty}k(k-1)(\varepsilon_k - \varepsilon_{k+1})t^{k-2}\,,\end{aligned} \tag{3.15}$$

for $t = 1$ and after replacing λ with $\mu - \varepsilon'$ results in

$$\mu_{(2)} = \mu^2 - \varepsilon'^2 + 2\sum_{k=1}^{\infty}(k-1)\varepsilon_k\,. \tag{3.16}$$

In a similar manner, it can be shown that the third factorial moment is given by

$$\mu_{(3)} = \mu^3 - 3\mu\varepsilon'^2 + 2\varepsilon'^3 + 6(\mu - \varepsilon')\sum_{k=1}^{\infty}(k-1)\varepsilon_k + \sum_{k=0}^{\infty}k(k-1)(k-2)(\varepsilon_k - \varepsilon_{k+1})\,. \tag{3.17}$$

Since the method described above becomes quite cumbersome and very time consuming for the computation of higher factorial moments, we suggest to apply the following recurrence formula being useful for the successive evaluation

$$\mu_{(r)} = \sum_{i=0}^{r}\binom{r}{i}(\mu - \varepsilon')^i \sum_{k=0}^{\infty}k_{(r-i)}(\varepsilon_k - \varepsilon_{k+1})\,, \quad r = 1, 2, \ldots \tag{3.18}$$

Here, $k_{(r-i)}$ denote the descending or falling factorial which is defined as

$$k_{(r-i)} = \prod_{j=0}^{r-i-1}(k-j) = \frac{k!}{(k-r+i)!}\,,$$

where $k_{(r-i)} = 0$ for $r - i > k$.
Analogously to the procedure discussed in Appendix B, the initial and central moments of the Fucks GPD, denoted by μ'_k and μ_k, respectively, can be determined from the factorial moments using relations (B.9) and (B.13). Hence, the second and third initial moments result in

$$\begin{aligned}\mu'_2 &= \mu^2 + \mu - \varepsilon'^2 + 2\sum_{k=1}^{\infty}(k-1)\varepsilon_k\,,\\ \mu'_3 &= \mu^3 + 3\mu^2 + \mu - 3\mu\varepsilon'^2 - 6\mu\varepsilon' + 2\varepsilon'^3 + 3\varepsilon'^2 - \varepsilon' + \\ &\quad + 6(\mu - \varepsilon')\sum_{k=1}^{\infty}k\varepsilon_k + \sum_{k=1}^{\infty}k^3(\varepsilon_k - \varepsilon_{k+1})\,,\end{aligned} \tag{3.19}$$

whence the third central moment can be written as

$$\mu_3 = \mu + 2\varepsilon'^3 - 3\varepsilon'^2 + 6(1 - \varepsilon')\sum_{k=1}^{\infty}(k-1)\varepsilon_k + \sum_{k=1}^{\infty}k(k-1)(k-2)(\varepsilon_k - \varepsilon_{k+1})\,. \tag{3.20}$$

3.3 A Generalization of the Fucks GPD

Only a few years after the Russian translation of the Fucks' (1956a) article, three Georgian scholars, Cercvadze, Čikoidze, and Gačečiladze (1959) applied Fucks' ideas to Georgian linguistic material, mainly concentrating on phoneme frequencies and word length frequencies (see also Zerzwadse, Tschikoidse, and Gatschetschiladse, 1962). Moreover, they introduced a further generalization of the Fucks' (3.7) model which has been discussed in subsequent papers by Gačečiladze, Cercvadze, and Čikoidze (1961), Bokučava and Gačečiladze (1965), and Gačečiladze and Cilosani (1971). Later on, Piotrowski, Bektaev, and Piotrowskaja (1985, p. 269) termed this generalization the Fucks-Gačečiladze distribution.

Starting from the assumption that the process of text generation is a stochastic process Gačečiladze et al. (1961, p. 5) derived the following formula, representing the sum of weighted binomial probabilities

$$\pi_{i|p,q,\varepsilon_k} = \sum_{k=0}^{n} (\varepsilon_k - \varepsilon_{k+1}) \binom{n-k}{i-k} p^{i-k} q^{i-k} [1-p+p(1-q)]^{n-i}\,, \; i=0,\ldots,n\,, \quad (3.21)$$

where the difference $(\varepsilon_k - \varepsilon_{k+1})$ represents the statistical weight of the system's status before the beginning of the distribution process. Note that $\sum_{k=0}^{n} (\varepsilon_k - \varepsilon_{k+1}) = 1$ as the sum of all the weights, hence $\varepsilon_0 = 1 + \varepsilon_{n+1}$. If we assume that $p \in (0,1]$, then distribution (3.21) has the following form

$$\pi_{i|q,\varepsilon_k} = \sum_{k=0}^{n} (\varepsilon_k - \varepsilon_{k+1}) \binom{n-k}{i-k} \int\limits_0^1 (pq)^{i-k} (1-pq)^{n-i} dp\,, \;\; i=0,\ldots,n\,. \quad (3.22)$$

In a similar manner, as already demonstrated for the mean of the Fucks-binomial distribution (see Section 3.2), it can be shown that after summing up the corresponding ε_k expressions, the mean of the distribution (3.22) is given by

$$\mu = \sum_{i=0}^{n} i\, \pi_{i|q,\varepsilon_k} = \frac{q}{2} \left(n - \sum_{k=1}^{n} k(\varepsilon_k - \varepsilon_{k+1}) \right) + \sum_{k=1}^{n} k(\varepsilon_k - \varepsilon_{k+1})\,. \quad (3.23)$$

This relation differs from those mentioned by Gačečiladze et al. (1961)[2]. However, whenever the number of components k is not limited, the two equations will coincide, since then $\sum_{k=1}^{\infty} k(\varepsilon_k - \varepsilon_{k+1}) = \sum_{k=1}^{\infty} \varepsilon_k = \varepsilon'$. Taking $n \to \infty$ and $q \to 0$ such that $\mu - \varepsilon' = q(n-\varepsilon')/2 = const$ and by substituting $p = (t+1)/2$, the probability (3.22) can be written in the following form

$$\pi_{i|q,\varepsilon_k} = \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \binom{n-k}{i-k} \left(\frac{\mu - \varepsilon'}{n - \varepsilon'} \right)^{i-k} \frac{1}{2} \int_{-1}^{1} (t+1)^{i-k} \left(1 - (t+1) \frac{\mu - \varepsilon'}{n - \varepsilon'} \right)^{n-i} dt.$$

Since the last factor in the integral part of the above equation converges to $e^{-(t+1)(\mu-\varepsilon')}$ as n increases without a limit, as a result we have

$$\int_{-1}^{1} (t+1)^{i-k} \left(1 - \frac{(t+1)(\mu-\varepsilon')}{n-\varepsilon'} \right)^{n-i} dt \to e^{-(\mu-\varepsilon')} \int_{-1}^{1} (t+1)^{i-k} e^{-t(\mu-\varepsilon')} dt\,.$$

[2] Instead of the sum $\sum_{k=1}^{n} k(\varepsilon_k - \varepsilon_{k+1})$ in (3.23) Gačečiladze et al. (1961) have $\sum_{k=1}^{n-1} \varepsilon_k$.

Hence, by replacing $\mu - \varepsilon'$ with λ, we obtain as the limiting distribution of (3.22) the *Fucks-Gačečiladze distribution*[3] with the probability mass function given by

$$\pi^{\text{FG}}_{i|\lambda,\varepsilon_k} = P(X = i) = \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \frac{\lambda^{i-k}}{(i-k)!}\, \varphi_k(\lambda, i)\,, \tag{3.24}$$

where function $\varphi_k(\lambda, i)$ denotes

$$\varphi_k(\lambda, i) = \frac{1}{2} \int_{-1}^{1} (t+1)^{i-k} e^{-\lambda t} dt\,. \tag{3.25}$$

Analogously to the case of the Fucks' GPD (3.7), the condition $\varepsilon_0 = 1 \geq \varepsilon_k \geq \varepsilon_{k+1}$, for $k \geq 1$ is again required while the individual weights $(\varepsilon_k - \varepsilon_{k+1})$ are additionally multiplied by the function $\varphi_k(\lambda, i)$, which depends on the parameter λ and the relevant class i. For $\varphi_k(\lambda, i) = 1$ distribution (3.24) simplifies to the Fucks' GPD (3.7), as its special case. Note that equality $\varphi_0(\lambda, i) = \varphi_k(\lambda, i+k)$ holds for any $k \geq i$, where $i \in \mathbb{N}_0$. Furthermore, we show that the function $\varphi_i(\lambda) = \varphi_0(\lambda, i)$ satisfies the following recurrent relation

$$\varphi_i(\lambda) = -\frac{2^{i-1} e^{-\lambda}}{\lambda} + \frac{i}{\lambda}\, \varphi_{i-1}(\lambda)\,, \;\; i = 1, 2, \ldots \tag{3.26}$$

where $\varphi_0(\lambda) = (e^{\lambda} - e^{-\lambda})/2\lambda$. Denoting by $i - k = l$ and substituting $z = (t+1)\lambda$ in equation (3.25) we can rewrite the function $\varphi_k(\lambda, i)$ as

$$\varphi_l(\lambda) = \frac{1}{2} \int_0^{2\lambda} \left(\frac{z}{\lambda}\right)^l e^{\lambda - z} \frac{dz}{\lambda} = \frac{e^{\lambda}}{2\lambda^{l+1}} \int_0^{2\lambda} z^l e^{-z} dz\,. \tag{3.27}$$

Transforming the expression $\varphi_{i-1}(\lambda)$ at the right-hand side of the formula (3.26) by the relation (3.27) and expressing additionally the term $(2\lambda)^i e^{-2\lambda}$ as an integral

$$(2\lambda)^i e^{-2\lambda} = \int_0^{2\lambda} \frac{\partial(z^i e^{-z})}{\partial z} dz = \int_0^{2\lambda} (i z^{i-1} e^{-z} - z^i e^{-z}) dz\,,$$

we have

$$\varphi_i(\lambda) = \frac{e^{\lambda}}{2\lambda^{i+1}} \left(\int_0^{2\lambda} i z^{i-1} e^{-z} dz - (2\lambda)^i e^{-2\lambda} \right) = \frac{e^{\lambda}}{2\lambda^{i+1}} \int_0^{2\lambda} z^i e^{-z} dz$$

which is $\varphi_i(\lambda)$ given in (3.27). Another way to express the function $\varphi_i(\lambda)$ would imply the use of the Gamma function, defined as

$$\Gamma(a) = \int_0^{\infty} z^{a-1} e^{-z} dz\,, \;\; a > 0\,.$$

[3] Unfortunately, neither Gačečiladze et al. (1961) nor other authors mentioned above do not explain how to get distribution (3.24), they only present this final formula. For this reason, we consider here its exact derivation.

Therefore, relation (3.27) can be written as

$$\varphi_i(\lambda) = \frac{e^{\lambda}}{2\lambda^{i+1}} \left(\int_0^{\infty} z^i e^{-z} dz - \int_{2\lambda}^{\infty} z^i e^{-z} dz \right) = \frac{e^{\lambda}}{2\lambda^{i+1}} \left(\Gamma(i+1) - \int_{2\lambda}^{\infty} z^i e^{-z} dz \right).$$

Accordingly, substituting $z = a + 2\lambda$ and applying the binomial theorem

$$(a + 2\lambda)^i = \sum_{j=0}^{i} \binom{i}{j} (2\lambda)^{i-j} a^j$$

we obtain (cf. Antić et al., 2005, p. 170)

$$\varphi_i(\lambda) = \frac{e^{\lambda}}{2\lambda^{i+1}} \left(\Gamma(i+1) - e^{-2\lambda} \sum_{j=0}^{i} \binom{i}{j} (2\lambda)^{i-j}\, \Gamma(j+1) \right). \tag{3.28}$$

3.3.1 Probability Generating Function and Moments

The pgf of distribution (3.24) can be found directly by applying definition (B.1), Appendix B. Consequently, we have

$$G(t) = \sum_{i=0}^{\infty} t^i \pi^{\mathrm{FG}}_{i|\lambda,\varepsilon_k} = \sum_{i=0}^{\infty} t^i e^{-\lambda} \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \frac{\lambda^{i-k}}{(i-k)!}\, \varphi_k(\lambda, i)\,.$$

By summing up corresponding $(\varepsilon_k - \varepsilon_{k+1})$ terms and considering the fact that based on relation (3.25) the identity $\varphi_0(\lambda, i) = \varphi_k(\lambda, i+k)$ holds for all $k \geq i$, $i \in \mathbb{N}_0$, the above pgf can be rewritten as

$$G(t) = e^{-\lambda} \left(\varphi_0(\lambda, 0) + t\lambda\varphi_0(\lambda, 1) + \frac{t^2\lambda^2}{2!}\varphi_0(\lambda, 2) + \ldots \right) \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) t^k\,,$$

which by expressing $\varphi_0(\lambda, k)$ in terms of $\varphi_0(\lambda, 0)$ and some calculations simplifies to

$$G(t) = e^{-\lambda} \sum_{k=0}^{\infty} t^k \left(\varphi_0(\lambda, 0) - e^{-\lambda} t \sum_{k=0}^{\infty} \frac{(2t\lambda)^k}{(k+1)!} \right) \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) t^k\,.$$

Ultimately, substitution of the series expansions and $\varphi_0(\lambda, 0) = (e^{\lambda} - e^{-\lambda})/2\lambda$ gives the pgf of distribution (3.24) in the form (cf. Gačečiladze et al., 1961, p. 6) [4]

$$G(t) = \frac{1}{2} \frac{e^{2\lambda(t-1)} - 1}{\lambda(t-1)} \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) t^k\,. \tag{3.29}$$

[4] This formula is presented by the authors, however the process of its derivation is not explained neither in any of their papers nor in other related contributions.

The expressions for the factorial moments can be evaluated from the formula (3.29) as described in Appendix B. Applying the relations (3.12) and (3.14), the first three factorial moments of the distribution (3.24) are

$$\begin{aligned}
\mu_{(1)} &= \varepsilon' + \lambda\,,\\
\mu_{(2)} &= \frac{4}{3}\lambda^2 + (2\lambda - 2)\varepsilon' + 2\sum_{k=1}^{\infty} k\varepsilon_k\,,\\
\mu_{(3)} &= 2\lambda^3 + 4\lambda^2\varepsilon' + (5 - 6\lambda)\varepsilon' + 6(\lambda - 1)\sum_{k=1}^{\infty} k\varepsilon_k + \sum_{k=0}^{\infty} k^3(\varepsilon_k - \varepsilon_{k+1})\,,
\end{aligned}$$

which by substituting $\lambda = \mu - \varepsilon'$ yields

$$\begin{aligned}
\mu_{(1)} &= \mu\,,\\
\mu_{(2)} &= \frac{4}{3}\mu^2 - \frac{2}{3}\mu\varepsilon' - \frac{2}{3}\varepsilon'^2 - 2\varepsilon' + 2\sum_{k=1}^{\infty} k\varepsilon_k\,,\\
\mu_{(3)} &= 2\mu^3 - 2\mu^2\varepsilon' - 2\mu\varepsilon'^2 - 6\mu\varepsilon' + 2\varepsilon'^3 + 6\varepsilon'^2 + 5\varepsilon' +\\
&\quad + 6(\mu - \varepsilon' - 1)\sum_{k=0}^{\infty} k\varepsilon_k + \sum_{k=0}^{\infty} k^3(\varepsilon_k - \varepsilon_{k+1})\,.
\end{aligned} \tag{3.30}$$

Following the derivation analogy approach of Fucks' GPD, initial moments of the distribution (3.24) can be derived as

$$\begin{aligned}
\mu'_2 &= \frac{4}{3}\mu^2 - \frac{2}{3}\mu\varepsilon' + \mu - \frac{2}{3}\varepsilon'^2 - 2\varepsilon' + 2\sum_{k=1}^{\infty} k\varepsilon_k\,,\\
\mu'_3 &= 2\mu^3 + 4\mu^2 - 2\mu^2\varepsilon' - 2\mu\varepsilon'^2 + \mu - 8\mu\varepsilon' + 2\varepsilon'^3 + 4\varepsilon'^2 - \varepsilon' +\\
&\quad + 6(\mu - \varepsilon')\sum_{k=1}^{\infty} k\varepsilon_k + \sum_{k=1}^{\infty} k^3(\varepsilon_k - \varepsilon_{k+1})\,.
\end{aligned} \tag{3.31}$$

whence the second and third central moments of distribution (3.24) result in

$$\begin{aligned}
\mu_2 &= \frac{1}{3}\mu^2 - \frac{2}{3}\mu\varepsilon' + \mu - \frac{2}{3}\varepsilon'^2 - 2\varepsilon' + 2\sum_{k=1}^{\infty} k\varepsilon_k\,,\\
\mu_3 &= \mu^2 + \mu - 2\mu\varepsilon' + 2\varepsilon'^3 + 4\varepsilon'^2 - \varepsilon' - 6\varepsilon'\sum_{k=1}^{\infty} k\varepsilon_k + \sum_{k=1}^{\infty} k^3(\varepsilon_k - \varepsilon_{k+1})\,.
\end{aligned} \tag{3.32}$$

3.4 Parameter Estimation

Let us assume that the probability mass function, given in (3.7), is known, except for its parameters ε_k and λ. We know the theoretical representation of the observed distribution if we find the estimated values of the unknown parameters.

3.4.1 Estimation by Method of Moments

To estimate the ε_k values, the simplest method to apply, also suggested by Fucks (1956a, 12f.), is the method of moments. Since the moments are represented as polynomials in ε_k, one obtains estimates of the unknown parameters by substituting the theoretical moments by the empirical ones and by solving algebraic equations for ε_k. The estimation process is rather complex because many equations with unknown parameters ε_k require their solution. Below, it will be shown that the estimation process is getting much easier for special cases of the Fucks' GPD. Also, we will show that the 1-displaced Poisson distribution is a special case of the Fucks' GPD for a particular choice of ε_k values. Two more special cases which can be derived from Fucks' GPD will be additionally discussed in detail. Depending on the number of parameters taken into consideration, one may distinguish one-, two-, and three-parameter special cases of Fucks' GPD, which will be considered in this order.

3.4.1.1 A One-Parameter Case of the Fucks' GPD

Let us first consider the one-parameter model as the simplest of all special cases mentioned. Here, only one parameter has to be estimated. Assuming that $\varepsilon_0 = 1$ and $\varepsilon_k = 0$ for $k \geq 1$, the Fucks' GPD (3.7) leads to the *Poisson distribution*

$$\pi^{\mathrm{P}}_{i|\lambda} = \frac{e^{-\lambda}\lambda^i}{i!}\,, \quad i = 0, 1, 2, \dots \tag{3.33}$$

where $\lambda = \mu > 0$ and $\mathrm{E}(X) = \mathrm{var}(X) = \mu$.
Choosing $\varepsilon_0 = \varepsilon_1 = 1$ and $\varepsilon_k = 0$, $k \geq 2$, i.e. assuming that our sample has no zero-syllable words, we obtain the *1-displaced Poisson distribution*

$$\pi^{\mathrm{1dP}}_{i|\lambda} = \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!}\,, \quad i = 1, 2, \dots \tag{3.34}$$

Here, $\lambda = \mu - 1 > 0$, $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \mu - 1$.

3.4.1.2 A Two-Parameter Case of the Fucks' GPD

Setting $\varepsilon_0 = 1$, $\varepsilon_1 = \alpha$ and $\varepsilon_k = 0$, $k \geq 2$ in Fucks' GPD (3.7), yields a two-parameter distribution. This distribution, termed also *Dacey-Poisson distribution* in contemporary research (cf. Wimmer and Altmann, 1999, p. 111), has been discussed by Fucks (1955, p. 207) as another special case of his GPD, though not by this name, and only in its 1-displaced form. In its ordinary setting, it takes the following form

$$\begin{aligned} \pi^{\mathrm{DP}}_{0|\lambda,\alpha} &= e^{-\lambda}(1-\alpha)\,, \\ \pi^{\mathrm{DP}}_{i|\lambda,\alpha} &= e^{-\lambda}\left((1-\alpha)\frac{\lambda^i}{i!} + \alpha\frac{\lambda^{i-1}}{(i-1)!}\right)\,, \quad i = 1, 2, \dots \end{aligned} \tag{3.35}$$

where $\lambda = \mu - \alpha > 0$, $0 \leq \alpha \leq 1$, $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \mu - \alpha^2$.
Hence, in addition to λ, a second parameter α has to be estimated. Notice that under

a particular condition, namely for $\mu = 2\alpha$, the Dacey-Poisson distribution yields the so-called *Kemp-Kemp-Poisson distribution* with parameter α (cf. Wimmer and Altmann, 1999, p. 344), being therefore a one-parameter special case of the Fucks' GPD. Its probability mass function is given by

$$\pi_{i|\alpha}^{\mathrm{KKP}} = \frac{e^{-\alpha}\alpha^{i}}{i!}(i - \alpha + 1)\,, \quad i = 0, 1, 2, \ldots \tag{3.36}$$

Here, $\lambda = \alpha$, $0 < \alpha \leq 1$, $\mathrm{E}(X) = 2\alpha$ and $\mathrm{var}(X) = \alpha(2 - \alpha)$.
Similarly, for $\varepsilon_0 = \varepsilon_1 = 1$, $\varepsilon_2 = \alpha$ and $\varepsilon_k = 0$, $k \geq 3$, the *1-displaced Dacey-Poisson* model results from the Fucks' GPD (3.7) (cf. Wimmer and Altmann, 1999, p. 111) as

$$\begin{aligned} \pi_{1|\lambda,\alpha}^{\mathrm{1dDP}} &= e^{-\lambda}(1-\alpha)\,, \\ \pi_{i|\lambda,\alpha}^{\mathrm{1dDP}} &= e^{-\lambda}\left((1-\alpha)\frac{\lambda^{i-1}}{(i-1)!} + \alpha\frac{\lambda^{i-2}}{(i-2)!}\right)\,, \quad i = 2, 3, \ldots\,, \end{aligned} \tag{3.37}$$

with $\lambda = \mu - \alpha - 1$, $0 \leq \alpha \leq 1$, $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \mu - 1 - \alpha^2$.

Now, using the first two moments of the given model, referring to (3.8) and (3.9), we can find the moment estimates of the parameters λ and α. These are summarized in Table 3.2 for the one- and two-parameter special case of the Fucks' GPD.

Table 3.2: Moment estimates of the special cases of the Fucks' GPD

Distribution	Restriction	Estimates
Poisson	$\varepsilon_0 = 1$, $\varepsilon_k = 0,\ k \geq 1$	$\hat{\lambda} = \bar{x}$
1-displaced Poisson	$\varepsilon_0 = \varepsilon_1 = 1$, $\varepsilon_k = 0,\ k \geq 2$	$\hat{\lambda} = \bar{x} - 1$
Kemp-Kemp-Poisson	$\mu = 2\alpha,\ \varepsilon_0 = 1$, $\varepsilon_1 = \alpha,\ \varepsilon_k = 0,\ k \geq 2$	$\hat{\alpha} = \bar{x}/2$
Dacey-Poisson	$\varepsilon_0 = 1,\ \varepsilon_1 = \alpha$, $\varepsilon_k = 0,\ k \geq 2$	$\hat{\alpha} = \sqrt{\bar{x} - s^2}$, $\hat{\lambda} = \bar{x} - \hat{\alpha}$
1-displaced Dacey-Poisson	$\varepsilon_0 = \varepsilon_1 = 1,\ \varepsilon_2 = \alpha$, $\varepsilon_k = 0,\ k \geq 3$	$\hat{\alpha} = \sqrt{\bar{x} - 1 - s^2}$, $\hat{\lambda} = \bar{x} - 1 - \hat{\alpha}$

3.4.1.3 A Three-Parameter Case of the Fucks' GPD

In case of the three-parameter model, in addition to λ, two more parameters from the whole ε-spectrum have to be estimated, namely ε_2 and ε_3. However, the estimation depends on whether a class of zero-syllable words has to be taken into consideration, or not. Here, we assume that our sample has no zero-syllable words.

Setting $\varepsilon_0 = \varepsilon_1 = 1$, $\varepsilon_2 = \alpha$, $\varepsilon_3 = \beta$, and $\varepsilon_k = 0$, $k \geq 4$, results in a model, here called *1-displaced three-parameter Fucks model*, with probability mass function given by

$$\begin{aligned}
\pi_{1|\lambda,\alpha} &= e^{-\lambda}(1-\alpha)\,,\\
\pi_{2|\lambda,\alpha,\beta} &= e^{-\lambda}\left((1-\alpha)\lambda + (\alpha-\beta)\right)\,,\\
\pi_{i|\lambda,\alpha,\beta} &= e^{-\lambda}\left((1-\alpha)\frac{\lambda^{i-1}}{(i-1)!} + (\alpha-\beta)\frac{\lambda^{i-2}}{(i-2)!} + \beta\frac{\lambda^{i-3}}{(i-3)!}\right)\,,\quad i \geq 3\,,
\end{aligned} \tag{3.38}$$

where $\lambda = \mu - 1 - \alpha - \beta > 0$ and condition $0 \leq \beta \leq \alpha \leq 1$ holds. The first two moments of distribution (3.38) are $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \mu - 1 - (\alpha+\beta)^2 + 2\beta$. In Section 3.5 we discuss restrictions of this model based on the variance to mean ratio. However, there seem to be also other limitations, discussed below, responsible for the possible inadequacy of this model.

In the next step, α and β have to be estimated using the second and third theoretical central moment of the Fucks' GPD. Substituting the above given ε-values in (3.9) and (3.20) we obtain the following system of equations

$$\begin{aligned}
\mu_2 &= \mu - (1+\alpha+\beta)^2 + 2(\alpha+2\beta) = \mu - 1 - (\alpha+\beta)^2 + 2\beta\,,\\
\mu_3 &= \mu + 2(1+\alpha+\beta)^3 - 3(1+\alpha+\beta)^2 - 6(\alpha+\beta)(\alpha+2\beta) + 6\beta\,.
\end{aligned} \tag{3.39}$$

By replacing $\alpha_+ = \alpha + \beta$ we get a simplified system

$$\begin{aligned}
\mu_2 &= \mu - 1 - \alpha_+^2 + 2\beta\,,\\
\mu_3 &= \mu - 1 + 2\alpha_+^3 - 3\alpha_+^2 - 6\alpha_+\beta + 6\beta\,,
\end{aligned} \tag{3.40}$$

whose solution is a cubic equation with regard to the parameter α_+ given by

$$\alpha_+^3 - 3\alpha_+(\mu - 1 - \mu_2) + \mu_3 - 3\mu_2 + 2\mu - 2 = 0\,. \tag{3.41}$$

There are three possible solutions to this equation, not all of which are necessarily real ones. For each real solution $a = \hat{\alpha}_+$ (possibly a_i, $i = 1, 2, 3$), the values for $\varepsilon_2 = \alpha$ and $\varepsilon_3 = \beta$ have to be estimated, what can quite easily be done by computer programs. Furthermore, estimated values have to fulfill condition (a), postulated by Fucks (see Section 3.2), which requires to satisfy inequality $\hat{\alpha} \geq \hat{\beta}$. Using the fact that $a = \hat{\alpha}_+ = \hat{\alpha} + \hat{\beta}$ and $s^2 = \bar{x} - 1 - a^2 + 2\hat{\beta}$ in inequality above, we have

$$\frac{2a - a^2 - s^2 + \bar{x} - 1}{2} \geq \frac{a^2 + s^2 - \bar{x} + 1}{2}\,, \tag{3.42}$$

which by defining the difference $M = \bar{x} - s^2$ can be simplified as

$$a^2 - a + 1 - M \leq 0\,.$$

As a consequence, one obtains the following two conditions:

(a) The sum $a = \hat{\alpha} + \hat{\beta}$ must be in a particular interval for each of the three possible solutions of a, i.e. $a_i \in [a_{i1}, a_{i2}]$ for $i = 1, 2, 3$, where

$$a_{i1} = \frac{1 - \sqrt{4M - 3}}{2} \qquad \text{and} \qquad a_{i2} = \frac{1 + \sqrt{4M - 3}}{2}\,.$$

(b) In order to have $a \in \mathbb{R}$, $4M - 3 \geq 0$ must hold, i.e. $M = \bar{x} - s^2 \geq 0.75\,.$

Summing up, the three-parameter Fucks' model is adequate only for particular types of empirical distributions, and it can not serve as an overall model for a language, not even for languages which form their words from syllables, as Fucks himself considered. One possible reason might be the specific parameter estimating procedure suggested by Fucks, which subsequently inspired some authors, though generally following Fucks' ideas, to find alternative ways to estimate the parameters of the Fucks' GPD.

3.4.1.4 A Three-Parameter Case of the Fucks-Gačečiladze Distribution

Opposite to Fucks' approach, Gačečiladze and Cilosani (1971, p. 115) suggested to estimate the unknown parameters of the Fucks-Gačečiladze model (3.24) not with recourse to the central moments, but by deriving the theoretical initial moments from the probability generating function (3.29). Obviously, central and initial moments can be transformed into each other, hence both methods lead to identical parameter estimates.

Setting $\varepsilon_0 = \varepsilon_1 = 1$ to exclude the class of zero-syllable words as well as $\varepsilon_2 = \alpha$, $\varepsilon_3 = \beta$, and $\varepsilon_k = 0$ for $k \geq 4$ results in the *1-displaced three-parameter Fucks-Gačečiladze distribution* with probability mass function given by

$$\begin{aligned}
\pi_{1|\lambda,\alpha} &= e^{-\lambda}(1-\alpha)\varphi_1(\lambda, 1)\,,\\
\pi_{2|\lambda,\alpha,\beta} &= e^{-\lambda}\left((1-\alpha)\lambda\varphi_1(\lambda,2) + (\alpha-\beta)\varphi_2(\lambda,2)\right)\,, \qquad (3.43)\\
\pi_{i|\lambda,\alpha,\beta} &= e^{-\lambda}\Bigg((1-\alpha)\frac{\lambda^{i-1}}{(i-1)!}\varphi_1(\lambda,i) + (\alpha-\beta)\frac{\lambda^{i-2}}{(i-2)!}\varphi_2(\lambda,i) + \beta\frac{\lambda^{i-3}}{(i-3)!}\varphi_3(\lambda,i)\Bigg), i \geq 3\,,
\end{aligned}$$

where $\lambda = \mu - 1 - \alpha - \beta > 0$ and $0 \leq \beta \leq \alpha \leq 1$. Here, the mean is $\mathrm{E}(X) = \mu$ and the variance is

$$\mathrm{var}(X) = \frac{1}{3}\Bigg(\mu^2 - 2\mu(\alpha+\beta-0.5) - 2(1+\alpha+\beta)^2 + 6(\alpha+2\beta)\Bigg). \qquad (3.44)$$

Denoting $a = 1 + \alpha + \beta$, equation system (3.31) simplifies to the following equations necessary for the estimation of the unknown parameters

$$\begin{aligned}
\mu_2' &= \frac{4}{3}\mu^2 - \frac{2}{3}\mu a + \mu - \frac{2}{3}a^2 + 2(\alpha+2\beta)\,,\\
\mu_3' &= 2\mu^3 - 2\mu^2 a + 4\mu^2 - 2\mu a^2 + \mu - 8\mu a + 2a^3 + 4a^2 + \qquad (3.45)\\
&\quad + 6(\mu - a)(a + \alpha + 2\beta) + 6\alpha + 18\beta\,.
\end{aligned}$$

3.4.2 Estimation by μ, μ_2 and First Frequency Class

An alternative way to estimate the two parameters ε_2 and ε_3 of the three-parameter Fucks' distribution (3.38) was suggested by two Polish authors, Bartkowiakowa and Gleichgewicht (1964, 1965). Their considerations are based on the assumption that each word can be divided into two syllable-parts. The first part, named *word beginning* can have one, two or maximally three syllables and is described by the random variable X_B with the range $\{1, 2, 3\}$. The subsequent part or *word end*, independent from the first one is supposed to be Poisson distributed with parameter λ, called here X_E. Furthermore, starting from the basis of a certain number of words, they assumed that the proportion of one-syllable word beginnings is $(1 - \varepsilon_2)$, the proportion of two-syllable word beginnings is $(\varepsilon_2 - \varepsilon_3)$ and that of three-syllable word beginnings is ε_3. Therefore, the distribution of the random variable $X = X_B + X_E$, denoting the number of syllables per word, is given by

$$\pi_{i|\lambda,\varepsilon_k} = P(X = i) = \sum_{j=1}^{3} P(X_B = j, X_E = i - j) \overset{ind.}{=} \sum_{j=1}^{3} P(X_B = j)P(X_E = i - j)\,,$$

wherefrom the following special cases are obtained

$$\begin{aligned} \pi_{1|\lambda,\varepsilon_2} &= (1 - \varepsilon_2)g_0\,, \\ \pi_{2|\lambda,\varepsilon_2,\varepsilon_3} &= (1 - \varepsilon_2)g_1 + (\varepsilon_2 - \varepsilon_3)g_0\,, \\ \pi_{i|\lambda,\varepsilon_2,\varepsilon_3} &= (1 - \varepsilon_2)g_{i-1} + (\varepsilon_2 - \varepsilon_3)g_{i-2} + \varepsilon_3 g_{i-3}\,, \qquad i \geq 3\,, \end{aligned} \tag{3.46}$$

in which g_k identifies the Poisson probabilities

$$g_k = P(X_E = k) = \frac{\lambda^k}{k!}e^{-\lambda}\,,\ \ k = 0, 1, 2, \ldots \tag{3.47}$$

Using the independency of the random variables X_B and X_E we have

$$\begin{aligned} \mu = \mathrm{E}(X) &= \mathrm{E}(X_B) + \mathrm{E}(X_E) = 1 + \varepsilon_2 + \varepsilon_3 + \lambda\,, \\ \mathrm{var}(X) &= \mathrm{var}(X_B) + \mathrm{var}(X_E) = \mu - (1 + \varepsilon_2 + \varepsilon_3)^2 + 2(\varepsilon_2 + 2\varepsilon_3)\,. \end{aligned} \tag{3.48}$$

It may be noted that distribution (3.46) is only a reformulated version of the three-parameter Fucks' distribution (3.38) using the Poisson distribution (3.47) with the parameter $\lambda = \mu - 1 - \varepsilon_2 - \varepsilon_3$.

As to the estimation of the unknown parameters ε_2 and ε_3, the authors did not apply the method of moments, as did Fucks, since results were not satisfactory. Their method is rather based on the adjustment of the distribution on the proportion of one-syllable words that were most frequent in their studied works (cf. Bartkowiakowa and Gleichgewicht, 1964, p. 347). Hence, using a logarithmic transformation of the proportion $\pi_{1|\lambda,\varepsilon_2}$ in formula (3.46), we obtain as the first equation

$$\log\left(\frac{\pi_{1|\lambda,\varepsilon_2}}{1 - \varepsilon_2}\right) = \log g_0\,.$$

Since $g_0 = e^{-\lambda}$ we have

$$\log\left(\frac{\pi_{1|\lambda,\varepsilon_2}}{1-\varepsilon_2}\right) = -\left(\mu - 1 - \varepsilon_2 - \varepsilon_3\right).$$

The second equation required for the system of equations is then gained from the variance specified in (3.48). Considering the empirical distribution, we get the following system of equations, adequate to arrive to a solution for ε_2 and ε_3

$$\begin{aligned} \log\left(\frac{p_1}{1-\hat{\varepsilon}_2}\right) &= -\left(\bar{x} - 1 - \hat{\varepsilon}_2 - \hat{\varepsilon}_3\right), \\ s^2 - \bar{x} &= -(1 + \hat{\varepsilon}_2 + \hat{\varepsilon}_3)^2 + 2(\hat{\varepsilon}_2 + 2\hat{\varepsilon}_3), \end{aligned} \tag{3.49}$$

where p_1 is the observed relative frequency of one-syllable words.

In addition to the different estimation procedure, the two Polish authors argued in favor of the χ^2-test in order to measure the goodness of fit of the theoretical distribution. Moreover, to improve the estimation results they suggested to apply the minimum χ^2 method, where the process of estimation is repeated as long as the minimal value for the χ^2 function is obtained.

3.5 Historical Applications and Limitations

In order to verify the accuracy of Fucks' statements we re-analyse his linguistic data given in Table 3.1. For this purpose we created nine artificial samples of an approximate size n=10000 each, by multiplying the given relative frequencies with 10000. Since the texts studied by Fucks contained no zero-syllable words[5], only the 1-displaced one-, two-, and three-parametric versions of the Fucks' GPD submodels are considered here.

Moreover, we studied under which empirical conditions some chosen model may be satisfactory for word length frequencies by applying the δ-value principle, introduced in Section 2.1. As can be seen from Table 3.3, the 1-displaced Poisson distribution (3.34) provides an adequate fit only for empirical samples with $d \approx 1$. In contrast, we have $\delta = 1 - \alpha^2/(\mu - 1)$ for the 1-displaced Dacey-Poisson model (3.37). The fact $\text{var}(X) > 0$ results in $\alpha^2/(\mu - 1) < 1$. Since $\lambda \geq 0$ as well as $0 \leq \alpha \leq 1$, we have $\alpha^2/(\mu - 1) = \alpha^2/(\lambda + \alpha) \geq 0$. Consequently $\delta \leq 1$, hence this two-parametric model may be applicable as a theoretical model for empirical samples having $d \leq 1$. Furthermore, for the Fucks' three-parametric model (3.38) the index of dispersion is given by

$$\delta = 1 - \frac{(\alpha + \beta)^2 - 2\beta}{\lambda + \alpha + \beta}. \tag{3.50}$$

Obviously, taking $\lambda \to \infty$, δ converges to unity. However, for $\lambda \to 0$ it is difficult to find a range for δ directly. For this reason, we create a grid of various parameter

[5] Note here Fucks' incorrect assumption that all words of nine languages mentioned in Table 3.1 are of at least one syllable. Actually, in Russian, *k, s, v* are zero-syllable words.

values with $0 \leq \beta \leq \alpha \leq 1$ and calculate the δ-value for each pair (α, β). Figure 3.3 provides the graphical illustration of our findings. The shaded triangle under the solid reference line $\alpha = \beta$ represents the domain of the parameters α and β. The blue colored area identifies those cases where $\delta \leq 1$ holds, whereas pairs with $\delta > 1$ are signified by red color. Table 3.4 exemplifies our obtained results for some certain parameter values. The pairs which are not allowed are marked by $\varnothing$. Evidently, choice $\alpha \in [0.5, 1]$ and $\beta \in [0, \alpha]$ implies $0 \leq \delta \leq 1$. As opposed to this, for $\beta \leq \alpha$, $\alpha \in [0, 0.5]$ and $\lambda \to 0$ results $0 \leq \delta < 2$ (cf. Table 3.5). Finally, we can conclude that the three-parametric Fucks' model is likely to be an appropriate one for empirical samples with $0 \leq d < 2$. In the case of the three-parametric Fucks-Gačečiladze model (3.43) the index of dispersion is getting even more complicated and is defined by

Table 3.3: Limitations for 1-displaced Fucks' sub-models based on index of dispersion

Distribution	E(X)	var(X)	$\delta = \dfrac{\text{var}(X)}{\text{E}(X) - 1}$
Poisson	μ	$\mu - 1$	$\delta = 1$
Dacey-Poisson	μ	$\mu - 1 - \alpha^2$	$\delta \leq 1$
3-param. Fucks	μ	$\mu - 1 - (\alpha + \beta)^2 + 2\beta$	$0 \leq \delta < 2$
3-param. Fucks-Gačečiladze	μ	see Formula (3.44)	$\delta \geq 0$

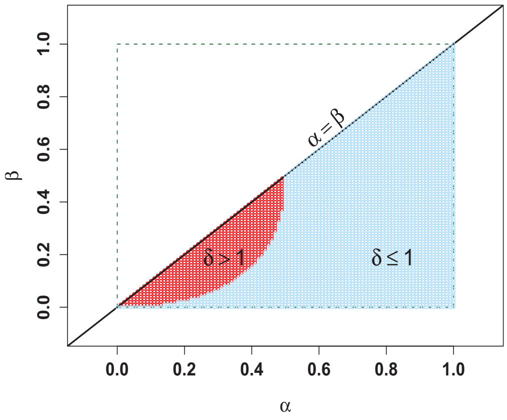


Figure 3.3: Under- and overdispersion area for the three-parametric Fucks' distribution

$$\delta = \frac{\lambda^2 + 3\lambda - 3\alpha^2 - 3\beta^2 + 3(\alpha + 3\beta - 2\alpha\beta)}{3(\lambda + \alpha + \beta)}\,.$$

Following the analogous approach as in the previous case we can conclude that for $\lambda \geq 2.5$ and $0 \leq \beta \leq \alpha \leq 1$, this distribution becomes over-dispersed ($\delta > 1$). For

Table 3.4: Three-parametric Fucks' model: index of dispersion δ for $\alpha \in [0.5, 1]$, $\beta \leq \alpha$, when $\lambda \to 0$

	α					
β	0.5	0.6	0.7	0.8	0.9	1
0	0.500	0.400	0.300	0.200	0.100	0.000
0.1	0.733	0.586	0.450	0.322	0.200	0.082
0.3	0.950	0.767	0.600	0.446	0.300	0.162
0.5	1.000	0.809	0.633	0.469	0.314	0.167
0.7	∅	∅	∅	0.433	0.275	0.124
0.9	∅	∅	∅	∅	0.200	0.047
1	∅	∅	∅	∅	∅	0.000

small values of λ this model seems to be a good choice for empirical samples meeting both conditions $0 \leq d \leq 1$ but also $d > 1$. The under- and overdispersion regions are identical with the corresponding areas of the three-parametric Fucks' model (cf. Figure 3.3).

Table 3.5: Three-parametric Fucks' model: index of dispersion δ for $0 \leq \beta \leq \alpha \leq 0.5$, when $\lambda \to 0$

	α									
β	0	0.05	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0	1	0.950	0.85	0.800	0.750	0.700	0.650	0.600	0.550	0.500
0.05	∅	1.899	1.30	1.150	1.033	0.936	0.850	0.772	0.700	0.632
0.15	∅	∅	1.70	1.507	1.350	1.217	1.100	0.995	0.900	0.812
0.25	∅	∅	∅	∅	1.500	1.359	1.233	1.119	1.014	0.917
0.35	∅	∅	∅	∅	∅	∅	1.300	1.183	1.075	0.974
0.45	∅	∅	∅	∅	∅	∅	∅	∅	1.100	0.997
0.5	∅	∅	∅	∅	∅	∅	∅	∅	∅	1.000

Further, we calculated the values of d for each of the nine observed languages. As shown in the Table 3.6, the value d ranges from 0.596 for Arabic up to 1.389 for English, indicating quite heterogeneous text samples. Additionally, Table 3.6 summarizes the results of the goodness of fit test, giving the standardized discrepancy index[6] C for each of the given models and languages where parameter estimates are evaluated by the method of moments.

It is obvious that the 1-displaced Poisson model does not equally fit linguistic data given by Fucks. The best fit is provided for Esperanto, good results are also achieved for Latin and German, but it seems to be an inadequate model for the

[6] For the explanation of the discrepancy index see Section 2.6.1.

remaining languages, especially English and Arabic. This result is also underlined by the values of d, since there is no good fit for those empirical samples where d significantly differs from unity. In some cases, the 1-displaced Dacey-Poisson model provides slightly better results than those obtained by 1-displaced Poisson distribution. Again, the best fit is obtained for Esperanto. Interestingly, it ensures a very good fit for Arabic and Turkish. Also for Latin, this model is useful. However, in cases denoted as $\varnothing$, no valid results can be obtained. This is due to the fact that the estimate of α, $\hat{\alpha} = \sqrt{\bar{x} - 1 - s^2}$, is not defined if $\bar{x} - 1 \leq s^2$. Checking the index of dispersion for Esperanto, Arabic, Latin, and Turkish it becomes clear, why the results for the two-parametric model are appropriate only for these languages. In all these cases $d < 1$ holds.

Fitting the three-parametric model to the Fucks' data, we observe good results in five of nine cases. Verification of the model is conformed by condition $0 \leq d < 2$ for each of the five samples. However, there are no solutions for the remaining four languages, although for all of them $d > 1$ is satisfied. Table 3.7 offers explanation for this findings, by specifying limitations responsible for the failure of Fucks' three-parameter model for English, German, Greek and Japanese. On the one hand, as soon as $M < 0.75$, the definition of the interval limits of a_1 and a_2 involves a negative argument for the root function. This is the case with the Japanese data. On the other hand, even if condition $M \geq 0.75$ is fulfilled, fitting Fucks' three-parameter model may fail, if $a \in [a_{i1}, a_{i2}]$ is not satisfied, as is in the case of English, German, and Greek.

Table 3.6: Results of fitting various 1-displaced models to Fucks' (1956a) data

Method of moments		**Discrepancy coefficient C**		
Language	**d**	**Poisson**	**Dacey-Poisson**	**3-param. Fucks**
English	1.3890	0.0816	$\varnothing$	$\varnothing$
German	1.1751	0.0168	$\varnothing$	$\varnothing$
Esperanto	0.9511	0.0025	0.0020	0.00003
Arabic	0.5964	0.1125	0.0083	0.0060
Greek	1.2179	0.0328	$\varnothing$	$\varnothing$
Japanese	1.2319	0.0333	$\varnothing$	$\varnothing$
Russian	1.1591	0.0208	$\varnothing$	0.0005
Latin	0.8704	0.0182	0.0148	0.0002
Turkish	0.8015	0.0232	0.0016	0.0016

Moreover, it is interesting to see how much the estimation procedure suggested by Bartkowiakowa and Gleichgewicht (1964) is able to improve the results of the fit. Table 3.8 compares the results obtained by two different parameter estimation methods: the original procedure as suggested by Fucks and the modification suggested by the Polish authors. The comparison is done only for those data sets, appropriate for the application of Fucks' three-parameter distribution model. Apparently, the approach of the Polish authors results in better estimates, at least for the data analyzed. A possible interpretation might have its justification in the fact that this estimation procedure is particularly adequate when the observed frequency in the first class is relatively large. Thus, more weight is given to this large frequency class, as compared to the the third moment which is more affected by the frequencies of the higher classes.

Table 3.7: Violations of the conditions for Fucks' three-parameter model

Method of moments	**English**	**German**	**Esperanto**	**Arabic**	**Greek**
C	∅	∅	<0.01	<0.01	∅
$\hat{\alpha}$	—	—	0.3933	0.7174	—
$\hat{\beta}$	—	—	0.0995	0.1805	—
$a = \hat{\alpha} + \hat{\beta}$	-0.0882	-0.1037	0.4929	0.8980	0.2800
a_{i1}	0.1968	0.1270	-0.0421	-0.3338	0.4108
a_{i2}	0.8032	0.8730	1.0421	1.3338	0.5892
$a_{i1} < a < a_{i2}$	–	–	✓	✓	–
$\bar{x}$	1.4064	1.6333	1.8971	2.1032	2.1106
s^2	0.5645	0.7442	0.8532	0.6579	1.3526
$M = \bar{x} - s^2$	0.8419	0.8891	1.0438	1.4453	0.7580
$M \geq 0.75$	✓	✓	✓	✓	✓
	Japanese	**Russian**	**Latin**	**Turkish**	
C	∅	< 0.01	< 0.01	< 0.01	
$\hat{\alpha}$	—	0.2083	0.5728	0.6164	
$\hat{\beta}$	—	0.1686	0.2416	0.1452	
$a = \hat{\alpha} + \hat{\beta}$	-0.1798	0.3769	0.8144	0.7616	
a_{i1}	$\mathbb{C}$	0.2659	-0.1558	-0.2346	
a_{i1}	$\mathbb{C}$	0.7341	1.1558	1.2346	
$a_{i1} < a < a_{i2}$	–	✓	✓	✓	
$\bar{x}$	2.1325	2.2268	2.3894	2.4588	
s^2	1.3952	1.4220	1.2093	1.1692	
$M = \bar{x} - s^2$	0.7374	0.8048	1.1800	1.2896	
$M \geq 0.75$	–	✓	✓	✓	

Finally, the conclusions of fitting the three-parametric Fucks–Gačečiladze model are given in Table 3.9. An acceptable result is now obtained for Greek, what was not the case when fitting the three-parameter Fucks' distribution (cf. Table 3.6). In fact, this generalization of Fucks' GPD provides a very good fit in six of the nine samples. Still, English, German and Japanese data can not be successfully modelled. The reason for this failure might be the fact that, for $\varphi_k(\lambda, i) = 1$, the

Table 3.8: Fucks' three-parameter model: two estimation methods

	Esperanto	**Arabic**	**Russian**	**Latin**	**Turkish**
Fucks (*Method of moments*)					
$\hat{\alpha}$	0.3933	0.7174	0.2083	0.5728	0.6164
$\hat{\beta}$	0.0995	0.1805	0.1686	0.2416	0.1452
C	0.00003	0.0060	0.0005	0.0002	0.0016
Polish (*Estimation by* μ, μ_2, π_1)					
$\hat{\alpha}$	0.3893	0.7148	0.2098	0.5744	0.6034
$\hat{\beta}$	0.0957	0.1599	0.1695	0.2490	0.1090
C	0.00001	0.0045	0.0005	0.0001	0.0012

Fucks–Gačečiladze distribution (3.24) simplifies to the Fucks' GPD, and under this condition provides the identical results.

Table 3.9: Three-parameter Fucks-Gačečiladze model: estimation by μ, μ_2, π_1

	English	**German**	**Esperanto**	**Arabic**	**Greek**
C	∅	∅	0.0008	0.0091	0.0148
$\hat{\alpha}$	—	—	0.4490	0.7251	0.3013
$\hat{\beta}$	—	—	0.1261	0.1986	0.1511
	Japanese	**Russian**	**Latin**	**Turkish**	
C	∅	0.0029	0.0034	0.0083	
$\hat{\alpha}$	—	0.3821	0.6230	0.6870	
$\hat{\beta}$	—	0.1885	0.3050	0.2606	

In summary, we can conclude that none of the discussed models above can be accepted as an overall valid theoretical model for word length frequencies, not even for syllabic languages, as Fucks himself claimed. Therefore, we search for the further possible generalizations of the Poisson distribution based on approaches discussed in Chapter 2.

Chapter 4

Singh-Poisson Distribution

4.1 Introduction

The Singh-Poisson (SP) distribution is a simple alternative to the Poisson distribution applicable in situations where the observed count data have $d \neq 1$, indicating that there is some deviation from the Poisson distribution. This distribution is a special case of a finite mixture[1], known also as *zero-modified Poisson distribution* where the Poisson distribution is combined with a one-point (degenerate) distribution concentrated at zero (cf. Johnson et al., 1992; Djuraš and Stadlober, 2010). It has two parameters denoted by α and θ and both of them are asked to be positive.

Let us remark that for $0 \leq \alpha \leq 1$ the SP distribution may also be derived as a Bernoulli mixture of Poisson distributions (cf. Wimmer and Altmann, 1999, p. 605). Starting from the conditional model $X|K \sim \text{Poisson}(K\theta)$ and assuming additionally parameter K to be randomized as $K \sim \text{Bernoulli}(\alpha)$, we obtain as a marginal model $X \sim \text{Singh-Poisson}(\alpha, \theta)$. Since the pgf of the Poisson$(K\theta)$ distribution satisfies the precondition (2.33) of Gurland's theorem 2.1 given in Section 2.2.1, we have

$$\text{Poisson}(K\theta) \bigwedge_K \text{Bernoulli}(\alpha) \sim \text{Bernoulli}(\alpha) \bigvee \text{Poisson}(\theta)$$

hence, the pgf of the mixed distribution of X has an alternative derivation as the pgf of the generalized Bernoulli distribution. Consequently, we have

$$G_X(t) = G_K(G_{X|K}(t|\theta)) = 1 - \alpha + \alpha e^{\theta(t-1)}\,. \tag{4.1}$$

Obviously, for $\alpha = 1$ it follows $G_X(t) = e^{\theta(t-1)}$ hence, SP distribution simplifies to an unmodified Poisson distribution with parameter θ. When $0 \leq \alpha \leq 1$ the model here is called the *zero-inflated Poisson distribution* (ZIP) or a *Poisson distribution with added zeroes*; see e.g. Cohen (1991) or Winkelmann (2000). In the last twenty years, much attention has been given to this model in the literature. Application areas include among others manufacturing defects (cf. Lambert, 1992), occupational

[1] For further details about finite mixtures see Section 2.2.1.

health injury data (cf. Lee, Wang, and Yau, 2001), dental epidemiology (cf. Dietz and Böhning, 2000) and zoological data, such as deaths of turtle juveniles caused by their exposure to the sun (cf. Özmen and Famoye, 2007). A good discussion on modeling count data having too many zeros, with particular focus on agricultural research, is available in Ridout, Demétrio, and Hinde (1998). Several applications of the ZIP model involving discrete data can also be found in Böhning (1998). Studied examples confirm that the ZIP distribution might be a useful alternative when the simple Poisson distribution does not provide an adequate fit due to an excess of zero counts in the observed data. To handle such overdispersed count data other authors proposed mixing a degenerate distribution with all mass at zero with distributions distinct from the Poisson. For example, Gupta, Gupta, and Tripathi (1996) suggested to combine it with Consul's generalized Poisson distribution[2] and used it to analyze overdispersed fetal lambs movement data and the numbers of death notices of older women appearing in the "*London Times*" on each day of three consecutive years.

However, for $1 < \alpha \leq 1/(1-e^{-\theta})$ we have a *zero-deflated Poisson distribution* (cf. e.g. Johnson et al., 1992; Dietz and Böhning, 2000) which can no longer appear as a mixture distribution. In practice, this case occurs seldom and thus has been much less explored. Recently, Kemp (2005) showed that the ability of this distribution to model underdispersion is very limited and complemented it by several illustrative examples. Moreover, the author compared it to the generalized Poisson underdispersed case which proved to have similar limitations. Obviously, substituting $\alpha = 1/(1-e^{-\theta})$ in relation (4.1) we get the pgf of the zero-truncated Poisson distribution. The SP model considered here incorporates both inflated and deflated case. In the next section we clarify that this distribution is both under- and overdispersed.

4.2 1-Displaced Singh-Poisson Distribution

In its 1-displaced form, the probability mass function of a discrete random variable X^d having SP distribution is given by

$$\pi^d_{x|\alpha,\theta} = P(X^d = x) = \begin{cases} 1-\alpha+\alpha e^{-\theta}, & x = 1 \\ \alpha\theta^{x-1}e^{-\theta}/(x-1)!, & x = 2,3,\ldots \end{cases} \tag{4.2}$$

where $\theta > 0$ and $0 \leq \alpha \leq \alpha_{\max} = 1/(1-e^{-\theta})$. Here, $\alpha_{\max}$ denotes the maximal possible value of α for given $\theta > 0$ and results from the inequality $1-\alpha+\alpha e^{-\theta} \geq 0$ which has to be satisfied. Moreover, note that for $\alpha = 1$ this distribution simplifies to the standard 1-displaced Poisson distribution with parameter θ.

The pgf of X^d can easily be derived using formula (2.36), Section 2.3.1 and

[2] This distribution was first introduced by Consul and Jain (1973a) and is studied in detail in Chapter 6.

formula (2.9), Section 2.2.1 or otherwise applying $G_X(t)$ from (4.1). Hence, we get

$$G_{X^d}(t) = tG_X(t) = t\left((1-\alpha) + \alpha e^{\theta(t-1)}\right) . \tag{4.3}$$

The mean and the variance of distribution (4.2) result from the above pgf as

$$\mu^d = \mathrm{E}(X^d) = 1 + \alpha\theta \quad \text{and} \quad \mathrm{var}(X^d) = \alpha\theta(1 + \theta - \alpha\theta) . \tag{4.4}$$

The factorial moments can be found by substituting equation (2.10), Section 2.2.1 into equation (2.38), Section 2.3.1. However, direct derivation from (4.3) is also possible, as explained in Appendix B. Consequently, we obtain

$$\mu^d_{(1)} = 1 + \alpha\theta \quad \text{and} \quad \mu^d_{(k)} = k\alpha\theta^{k-1} + \alpha\theta^k , \text{ for } k = 2, 3, \ldots \tag{4.5}$$

Table 4.1 summarizes the basic features of the distribution above in relation to those of its original and size-biased (see Section 4.3) versions.

Table 4.1: Singh-Poisson distribution: original, 1-displaced and size-biased forms

		Random Variable	
	X	X^d	X^*
Notation	SP(α,θ)	1+SP(α,θ)	1+P(θ)
Range	$\mathbb{N}_0$	$\mathbb{N}$	$\mathbb{N}$
pmf	$\pi_{x\|\alpha,\theta}$	$\pi_{x-1\|\alpha,\theta}$	$\dfrac{e^{-\theta}\theta^{x-1}}{(x-1)!}$
pgf	$1-\alpha+\alpha e^{\theta(t-1)}$	$tG_X(t)$	$te^{\theta(t-1)}$
E$(\cdot)$	$\alpha\theta$	$1+\alpha\theta$	$1+\theta$
var$(\cdot)$	$\alpha\theta(1+\theta-\alpha\theta)$	$\alpha\theta(1+\theta-\alpha\theta)$	θ

In order to get a better understanding of the behavior of the SP distribution (4.2) we plotted in Figures 4.1 and 4.2 a set of graphs for various values of parameters α and θ. Additionally, we compare SP probabilities to those of the 1-displaced Poisson with the same value of θ, displayed in the charts as blue dashed bar plots. The successive SP probabilities for possible x values are computed using the following ratio of recursive probabilities

$$\frac{\pi^d_{x|\alpha,\theta}}{\pi^d_{x-1|\alpha,\theta}} = \frac{\theta}{x-1} , \text{ for } x = 3, 4, \ldots , \tag{4.6}$$

where $\pi^d_{1|\alpha,\theta} = 1-\alpha+\alpha e^{-\theta}$ and $\pi^d_{2|\alpha,\theta} = \alpha\theta e^{-\theta}$. Each row in the figures corresponds to a different value of parameter α, whereas the columns show the effect of increasing parameter θ. Since the columns in both figures illustrate bar diagrams for two particular values of θ, namely $\theta = 0.8$ (left) and $\theta = 2.2$ (right), and differ in the value of α only, they clearly indicate the differences in the probability distributions that follow from modifications in the value of α. As long as $0 < \alpha < 1$, the probability $P(X^d = 1)$ is greater than $e^{-\theta}$, the 1-displaced Poisson probability at one and hence we have an excess of ones compared to the parent 1-displaced Poisson distribution. The closer α is to zero, the higher the proportion of ones is, while for $\alpha = 0$ we have a

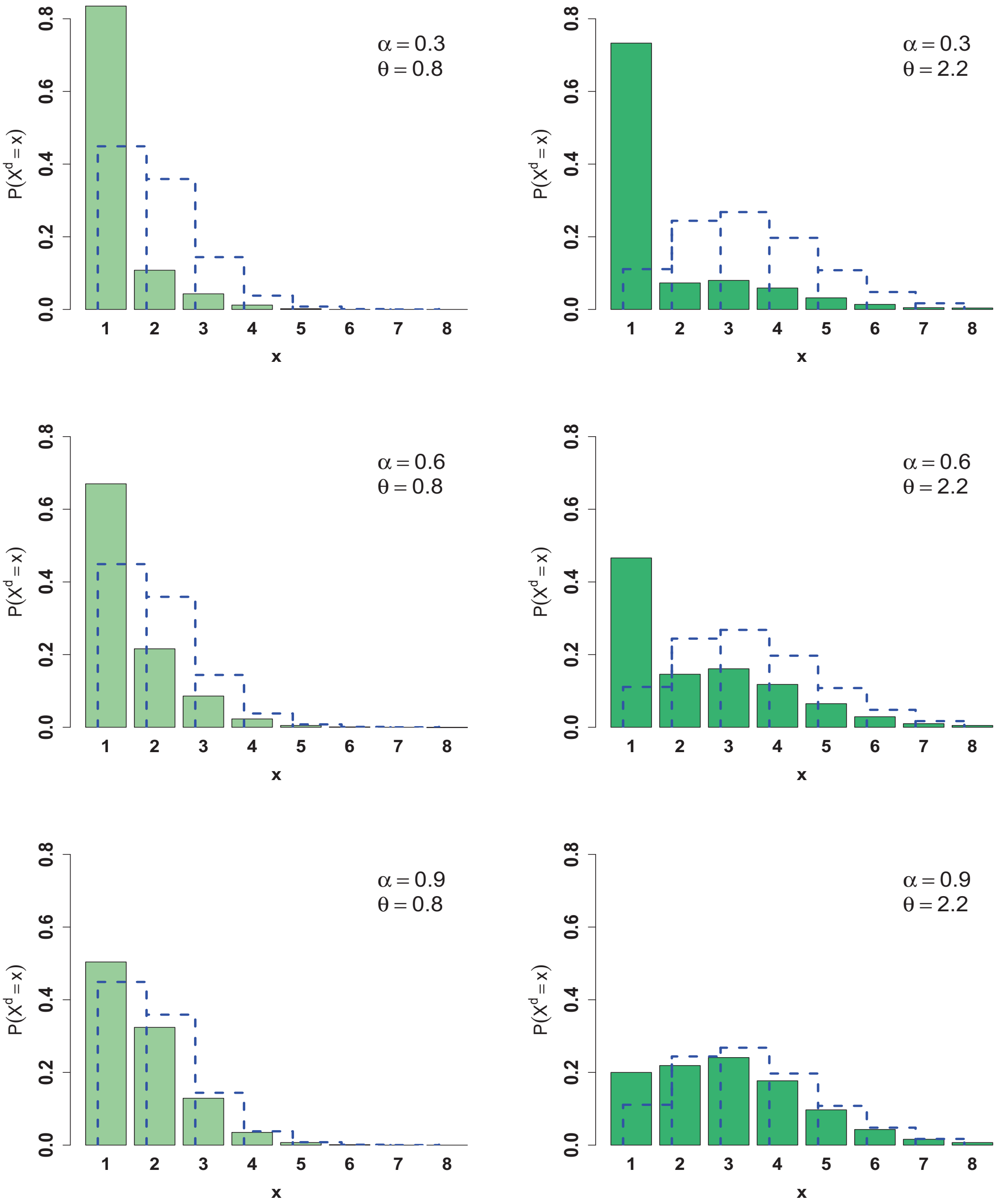


Figure 4.1: Graphs of probability distributions for $\theta = 0.8$ (left column), $\theta = 2.2$ (right column) and $\alpha = 0.3, 0.6, 0.9$ for the SP(α, θ) and the Poisson(θ) distribution.

degenerate, one-point distribution with all its mass at one. The *one-inflation* further implicates a reduction of the remaining frequencies by a corresponding amount (see Figure 4.1). The proportion of ones decreases as α converges to 1, and the SP probabilities are much closer to that of the Poisson. The total matching of the two distributions is given for $\alpha = 1$, regardless of the value of θ (see Figure 4.2, first row). As soon as $1 < \alpha \leq \alpha_{\max}$ we have *one-deflation* in SP compared to the Poisson

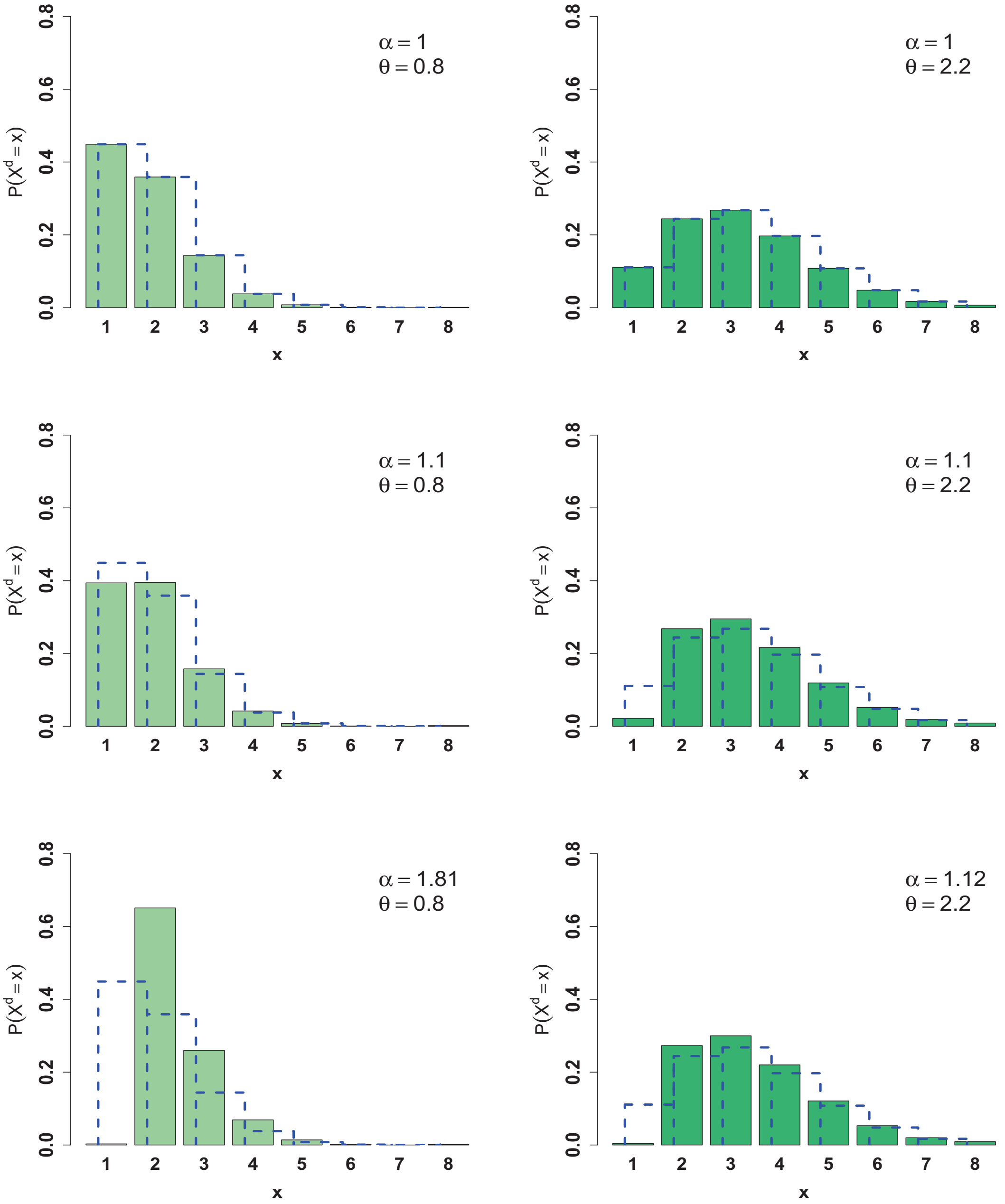


Figure 4.2: Graphs of probability distributions for $\theta = 0.8$ (left column), $\theta = 2.2$ (right column) and $\alpha = 1, 1.1, \alpha_{\max}$ for the SP(α, θ) and the Poisson(θ) distribution.

model, which influences also the remaining probability classes. The last row of the Figure 4.2 points out that the probability at one is reduced to zero when α reaches $\alpha_{\max}$, whereas values of α higher than this would give negative probability at one. It is also visible from the graphs that as θ becomes bigger, given a fixed value of α, the SP distribution requires a bigger interval on the x axis by spreading to the right side. As a consequence, probabilities in the first frequency class decrease.

An important characteristic of the 1-displaced SP distribution is its ability to model under- and overdispersion. Table 4.2 demonstrates both situations for different values of parameters α and θ. An invalid pair of parameters is denoted here by $\varnothing$. Such a case arises when for given θ the value of α exceeds $\alpha_{\max}$. The parameter α measures dispersion and hence tunes the type of the distribution. Justification is based on the following index of dispersion[3]

$$\delta = \frac{\mathrm{var}(X^d)}{\mathrm{E}(X^d) - 1} = \frac{\alpha\theta(\theta + 1 - \alpha\theta)}{\alpha\theta} = 1 + \theta(1 - \alpha)\,. \tag{4.7}$$

Clearly, under- or overdispersion is governed only by parameter α, as θ is positive. For $\alpha = 1$ we have equidispersion, i.e. $\delta = 1$. When $0 < \alpha < 1$, δ is strictly greater then 1 and we have overdispersion with respect to Poisson variation. The degree of overdispersion decreases as α increases to unity, also evident in the proportion of ones, which are now much closer to the Poisson probabilities (see Figure 4.1). However, for $\alpha = 0$ the variance becomes zero while the mean becomes unity, hence δ is indefinite. Therefore, we conclude that the 1-displaced SP distribution enables

Table 4.2: Over- and underdispersion in the 1-displaced Singh-Poisson distribution

		θ						
α		0.1	0.5	0.8	1	2.2	5	8
	μ^d	1.03	1.15	1.24	1.30	1.66	2.50	3.40
0.3	δ	1.07	1.35	1.56	1.70	2.54	4.50	6.60
	μ^d	1.06	1.30	1.48	1.60	2.32	4.00	5.80
0.6	δ	1.04	1.20	1.32	1.40	1.88	3.00	4.20
	μ^d	1.09	1.45	1.72	1.90	2.98	5.50	8.20
0.9	δ	1.01	1.05	1.08	1.10	1.22	1.50	1.80
	μ^d	1.10	1.50	1.80	2.00	3.20	6.00	9.00
1	δ	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	μ^d	1.11	1.55	1.88	2.10	3.42	$\varnothing$	$\varnothing$
1.1	δ	0.99	0.95	0.92	0.90	0.78	$\varnothing$	$\varnothing$
	μ^d	1.15	1.75	2.20	2.50	$\varnothing$	$\varnothing$	$\varnothing$
1.5	δ	0.95	0.75	0.60	0.50	$\varnothing$	$\varnothing$	$\varnothing$
	$\mu^d_{\max}$	2.051	2.271	2.453	2.582	3.474	6.034	9.003
$\alpha_{\max}$	$\delta_{\min}$	0.049	0.229	0.347	0.418	0.726	0.966	0.997

[3] For the explanation of this topic see Section 2.1.

us to model overdispersion, provided that $1 < \delta < 1 + \theta$. Notice also that for α fixed at any given value, the increase in the value of θ increases the mean as well as the variance (not explicitly shown in Table 4.2) of the SP distribution. When α increases from 1 to $\alpha_{\max}$ the distribution becomes underdispersed. For $\alpha = \alpha_{\max}$, we have $\delta = 1 - \theta/(e^{\theta} - 1) < 1$, which is obtained when $P(X^d = 1) = 0$, i.e. when the first probability class disappears (see Figure 4.2, last row). Notice that for the fixed value of α the degree of underdispersion decreases with increasing θ and ultimately reaches its minimum for $\alpha = \alpha_{\max}$. Therefore, as already recognized by Kemp (2005), the ability of the SP distribution to model underdispersion is limited. However, this limit need not be reached in practical situations.

4.3 Size-Biased Singh-Poisson Distribution

Following the concept of the size-biased distributions introduced in Section 2.3.2 and applying transformation (2.41) to the SP random variable X we obtain its size-biased version X^* with pmf given by

$$\pi^*_{x|\theta} = P(X^* = x) = \frac{\theta^{x-1} e^{-\theta}}{(x-1)!}\,, \quad \text{for } x = 1, 2, \dots \tag{4.8}$$

Notice that the parameter α disappears in the size-biased version of the SP model taking as such a very simple form, the well-known 1-displaced Poisson distribution with single parameter θ. The simplicity of its pgf $G_{X^*}(t) = te^{\theta(t-1)}$ enables easy computation of its factorial moments. The k-th factorial moment results in

$$\mu^*_{(k)} = k\theta^{k-1} + \theta^k\,, \quad \text{for } k = 1, 2, \dots \tag{4.9}$$

After simple calculations the mean and variance of X^* are given by

$$\mathrm{E}(X^*) = \mu^* = \theta + 1\,, \quad \mathrm{var}(X^*) = \theta\,, \tag{4.10}$$

whence the index of dispersion results in

$$\delta = \frac{\mathrm{var}(X^*)}{\mathrm{E}(X^*) - 1} = 1\,. \tag{4.11}$$

Clearly, distribution (4.8) is characterized by equidispersion.

4.4 Parameter Estimation

In this section we derive parameter estimators based on the three most common methods introduced in Section 2.4. As already mentioned above, the size-biased SP distribution is determined through a single parameter θ. The moment and maximum likelihood estimators of θ are identical and given by $\hat{\theta}_{\mathrm{MM}} = \hat{\theta}_{\mathrm{ML}} = \bar{x} - 1$. Therefore, parameter estimators obtained in the next sections refer to the 1-displaced SP model (4.2) only.

4.4.1 Estimation by Method of Moments

Moment estimators are obtained by equating $\mu = 1 + \alpha\theta$, the theoretical mean and $\mu_{(2)} = 2\alpha\theta + \alpha\theta^2$, the second theoretical factorial moment of the distribution (4.2), to their empirical counterparts $\bar{x}$ and $m_{(2)}$, respectively. Thus by solving simultaneously the following system of equations

$$\bar{x} = 1 + \alpha\theta \,, \qquad m_{(2)} = 2\alpha\theta + \alpha\theta^2 \,, \tag{4.12}$$

one gets simple moment (MM) estimators of the parameters α and θ as

$$\hat{\theta}_{\text{MM}} = \frac{m_{(2)}}{\bar{x} - 1} - 2 \text{ and } \hat{\alpha}_{\text{MM}} = \frac{\bar{x} - 1}{\hat{\theta}_{\text{MM}}} \,. \tag{4.13}$$

4.4.2 Estimation by Maximum Likelihood

Consider a random sample of size n from a distribution with pmf defined in (4.2). Let f_i denote the observed frequency of the i-th class such that $\sum_{i=1}^{k} f_i = n$, where k is the largest frequency class. The maximum likelihood (ML) estimator of parameter vector $\boldsymbol{\Theta} = (\alpha, \theta)$ is the value that maximizes the following likelihood function

$$L(\alpha, \theta | f_1, \ldots, f_k) = (1 - \alpha + \alpha e^{-\theta})^{f_1} \prod_{i=2}^{k} \left(\frac{\alpha\theta^{i-1} e^{-\theta}}{(i-1)!} \right)^{f_i} \,. \tag{4.14}$$

Therefore, the log-likelihood function is given by

$$\begin{aligned} & l(\alpha, \theta | f_1, \ldots, f_k) = \log L(\alpha, \theta | f_1, \ldots, f_k) = \\ & = f_1 \log\left(1 - \alpha + \alpha e^{-\theta}\right) + \sum_{i=2}^{k} f_i \left(\log \alpha + (i-1)\log\theta - \theta - \log(i-1)!\right) . \end{aligned} \tag{4.15}$$

The score equations are obtained by equating the first partial derivatives of (4.15) with respect to the parameters α and θ to zero. These equations are

$$\frac{\partial l(\alpha, \theta | f_1, \ldots, f_k)}{\partial \alpha} = \frac{(e^{-\theta} - 1) f_1}{1 - \alpha + \alpha e^{-\theta}} + \frac{1}{\alpha} \sum_{i=2}^{k} f_i = 0 \,, \tag{4.16}$$

$$\frac{\partial l(\alpha, \theta | f_1, \ldots, f_k)}{\partial \theta} = \frac{-\alpha e^{-\theta} f_1}{1 - \alpha + \alpha e^{-\theta}} + \frac{1}{\theta} \sum_{i=2}^{k} f_i (i - 1 - \theta) = 0 \,. \tag{4.17}$$

By solving equation (4.16) for α, the resulting ML estimator $\hat{\alpha}_{\text{ML}}$ of parameter α is

$$\hat{\alpha}_{\text{ML}} = \frac{n - f_1}{n(1 - e^{-\hat{\theta}_{\text{ML}}})} \,. \tag{4.18}$$

Multiplying equation (4.16) by α and adding it to (4.17) we obtain after a few algebraic simplifications the ML estimator $\hat{\theta}_{\text{ML}}$ of parameter θ as a solution of the transcendental equation

$$\frac{\theta(n - f_1)}{n(\bar{x} - 1)} + e^{-\theta} - 1 = 0 \,. \tag{4.19}$$

This equation can be solved for θ using function `uniroot()` available in R software.

4.4.3 Estimation Based on Mean and First Frequency Class

By equating the relative frequency of the first class f_1/n and the sample mean $\bar{x}$ to the probability of the first class $\pi^d_{1|\alpha,\theta} = 1 - \alpha + \alpha e^{-\theta}$ and the population mean $\mu = 1 + \alpha\theta$, respectively, we get the estimators of the parameters α and θ, which we denote here by $\hat{\alpha}_{\text{FF}}$ and $\hat{\theta}_{\text{FF}}$. After some calculations, it can be shown that these estimators are identical to the ML estimators $\hat{\alpha}_{\text{ML}}$ and $\hat{\theta}_{\text{ML}}$ obtained in Section 4.4.2.

4.5 A Simulation Study

To evaluate the performance of the above processed estimation procedures we made a simulation study where all three situations (i) overdispersion ($\delta > 1$), (ii) equidispersion ($\delta = 1$) and (iii) underdispersion ($\delta < 1$) were taken into account. Referring to the fact that the journalistic texts are overdispersed, the majority of the private letters and prose texts are equidispersed, whereas poems are underdispersed[4], we choose as model parameters (i) $(\alpha, \theta) = (0.82, 1.58)$, (ii) $(\alpha, \theta) = (0.92, 0.91)$ and (iii) $(\alpha, \theta) = (1.14, 0.63)$. These parameter settings coincide with the ML estimates of each text aggregation for Slovenian journalistic texts, private letters (i.e. prose), and poems, respectively, in order to get *representative texts* of each text type. For each of the three dispersion situations $M = 500$ Monte Carlo samples of size $n = 500$ and $n = 1000$ are drawn. To generate SP random variables we apply the inversion method introduced in Section 2.5, where the probabilities of the SP distribution are computed using recurrence formula (4.6), Section 4.2. The whole procedure was implemented by the software R, as already mentioned in Section 2.7.

The results of the simulation study are summarized in Table 4.3 for both sample sizes, namely $n = 500$ and $n = 1000$. For each of the three data situations MM and ML estimates have been calculated. The corresponding mean values of $M = 500$ estimated parameters $\hat{\alpha}$ and $\hat{\theta}$ are displayed in the second and fourth columns of Table 4.3. Additionally, we calculated the estimated standard errors of the mean values $\bar{\alpha}$ and $\bar{\theta}$ as the standard deviation of the $M = 500$ parameter estimates. These are labelled by $\text{se}_{\bar{\alpha}}$ and $\text{se}_{\bar{\theta}}$ and provide the precision of the resulting estimates (see third and fifth columns of Table 4.3). For both sample sizes, we observed similar results, independently of the data case and the estimation procedure applied. However, the standard errors of the estimated parameters are smaller for ML estimates by a factor of 0.90 or less compared to MM estimates and decrease with increasing sample size.

Figure 4.3 contains six plots visualizing the dependence between parameter estimates $\hat{\alpha}$ and $\hat{\theta}$ for $M = 500$ generated 1-displaced SP samples of size $n = 1000$ for both estimation methods and each of the three dispersion cases: (i) overdispersion (left column) (ii) equidispersion (middle column) and (iii) underdispersion (right column). Notice that the intensity of their correlation increases from overdispersed to underdispersed case, irrespective of the estimation procedure applied. The esti-

[4] This is the result of the statistical analysis of the Slovenian texts under study, discussed in more detail in Chapter 8.

Table 4.3: Estimation results for over-, equi- and underdispersed data situations

	$(\alpha,\theta)=(0.82,1.58)$			
$\delta>1$	$(\bar{\alpha}_{\text{MM}};\,\bar{\theta}_{\text{MM}})$	$(\text{se}_{\bar{\alpha}_{\text{MM}}};\,\text{se}_{\bar{\theta}_{\text{MM}}})$	$(\bar{\alpha}_{\text{ML}};\,\bar{\theta}_{\text{ML}})$	$(\text{se}_{\bar{\alpha}_{\text{ML}}};\,\text{se}_{\bar{\theta}_{\text{ML}}})$
$n=500$	$(0.823;1.579)$	$(0.037;0.093)$	$(0.821;1.582)$	$(0.030;0.081)$
$n=1000$	$(0.821;1.581)$	$(0.026;0.065)$	$(0.820;1.581)$	$(0.021;0.056)$
	$(\alpha,\theta)=(0.92,0.91)$			
$\delta=1$	$(\bar{\alpha}_{\text{MM}};\,\bar{\theta}_{\text{MM}})$	$(\text{se}_{\bar{\alpha}_{\text{MM}}};\,\text{se}_{\bar{\theta}_{\text{MM}}})$	$(\bar{\alpha}_{\text{ML}};\,\bar{\theta}_{\text{ML}})$	$(\text{se}_{\bar{\alpha}_{\text{ML}}};\,\text{se}_{\bar{\theta}_{\text{ML}}})$
$n=500$	$(0.928;0.907)$	$(0.066;0.079)$	$(0.925;0.909)$	$(0.057;0.071)$
$n=1000$	$(0.923;0.909)$	$(0.046;0.055)$	$(0.922;0.910)$	$(0.039;0.049)$
	$(\alpha,\theta)=(1.14,0.63)$			
$\delta<1$	$(\bar{\alpha}_{\text{MM}};\,\bar{\theta}_{\text{MM}})$	$(\text{se}_{\bar{\alpha}_{\text{MM}}};\,\text{se}_{\bar{\theta}_{\text{MM}}})$	$(\bar{\alpha}_{\text{ML}};\,\bar{\theta}_{\text{ML}})$	$(\text{se}_{\bar{\alpha}_{\text{ML}}};\,\text{se}_{\bar{\theta}_{\text{ML}}})$
$n=500$	$(1.156;0.628)$	$(0.107;0.068)$	$(1.152;0.629)$	$(0.097;0.063)$
$n=1000$	$(1.145;0.630)$	$(0.075;0.048)$	$(1.143;0.631)$	$(0.068;0.044)$

mated Pearson's correlation coefficient calculated as $\widehat{\rho}=cov(\hat{\alpha},\hat{\theta})/\text{se}_{\bar{\alpha}}\text{se}_{\bar{\theta}}$ measures the strength of this linear dependence and reconfirm the results obtained.

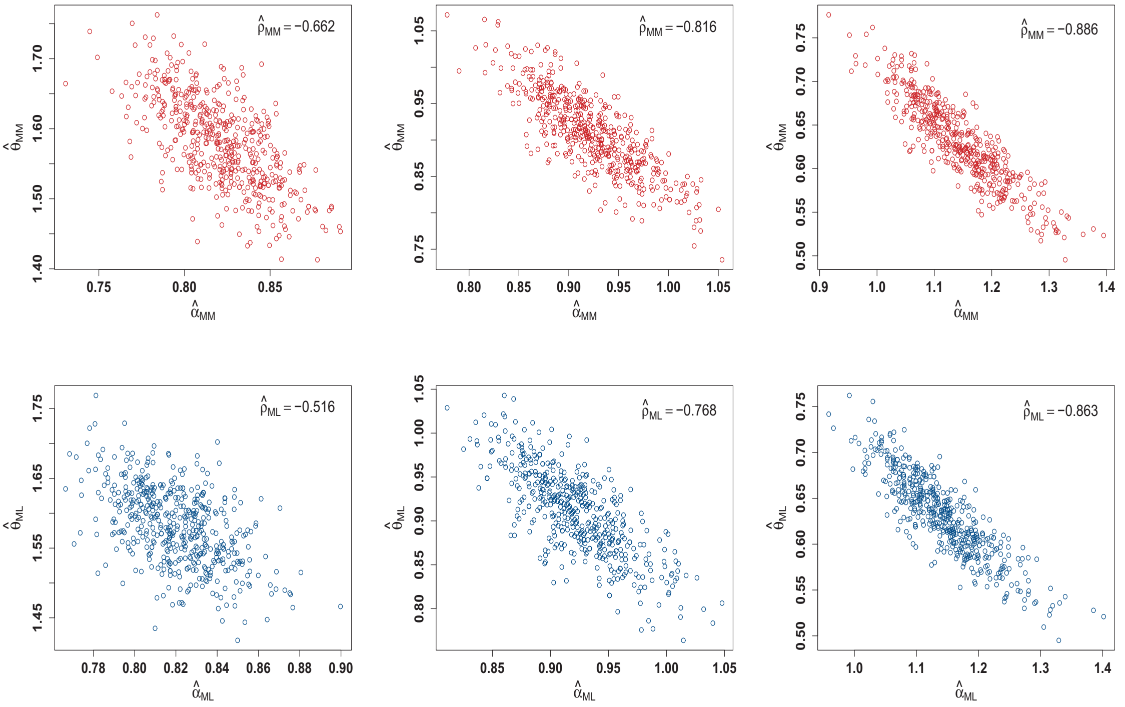


Figure 4.3: Scatterplots of the estimated parameters obtained from $M=500$ simulated 1-displaced SP samples of size $n=1000$ for over- (left column), equi- (middle column) and underdispersed (right column) data situation and both estimation methods applied.

Chapter 5

Hyper-Poisson Distribution

5.1 Introduction

The hyper-Poisson (HP) distribution is probably the most frequently used model of word length distributions, although it has also been mentioned in the literature as the model of the distribution of sentence length (cf. Antić et al., 2006b; Best, 2001; Kelih and Grzybek, 2004). This distribution is a two-parametric generalization of the Poisson distribution, and contains the Poisson distribution as a one-parameter subclass. To derive it, consider the conditional Poisson model $X|Y \sim \text{Poisson}(\theta Y)$ and assume additionally parameter Y to have a truncated Pearson Type III distribution with probability density function given by

$$g(y) = \frac{(\lambda-1)e^{\theta y}(1-y)^{\lambda-2}}{{}_1F_1[1;\lambda;\theta]}\,, \quad 0 \le y \le 1\,, \tag{5.1}$$

where $\lambda > 1$, $\theta > 0$ and ${}_1F_1[1;\lambda;\theta]$ denotes the *confluent hypergeometric function* (i.e. Kummer's function) with first argument equal to 1. As a result, the pmf of the mixture distribution follows as

$$\begin{aligned}\pi_{x|\lambda,\theta} = P(X=x) &= \int_0^1 \frac{e^{-\theta y}(\theta y)^x}{x!}\frac{(\lambda-1)e^{\theta y}(1-y)^{\lambda-2}}{{}_1F_1[1;\lambda;\theta]}\,dy \\ &= \frac{\theta^x(\lambda-1)}{x!\,{}_1F_1[1;\lambda;\theta]}\int_0^1 y^x(1-y)^{\lambda-2}dy = \frac{\theta^x\Gamma(\lambda)}{{}_1F_1[1;\lambda;\theta]\Gamma(\lambda+x)}\,, \quad x=0,1,\ldots\end{aligned} \tag{5.2}$$

where $\lambda > 0$ and $\theta > 0$, hence we obtain as a marginal model the HP distribution (cf. Bardwell and Crow, 1964). Some details are also given in Johnson et al. (1992, 193f). In its series representation the function ${}_1F_1[1;\lambda;\theta]$ has the form

$${}_1F_1[1;\lambda;\theta] = 1 + \frac{\theta}{\lambda} + \frac{\theta^2}{\lambda(\lambda+1)} + \ldots = \sum_{x=0}^{\infty}\frac{\theta^x}{\lambda^{(x)}}\,, \tag{5.3}$$

where $\lambda^{(x)} = \lambda(\lambda+1)\ldots(\lambda+x-1)$ is *Pochhammer's symbol*, also known as ascending factorial. Since $\lambda^{(x)} = \Gamma(\lambda+x)/\Gamma(\lambda)$ the pmf (5.2) can be rewritten as

$$\pi_{x|\lambda,\theta} = \frac{\theta^x}{{}_1F_1[1;\lambda;\theta]\lambda^{(x)}}\,, \quad x = 0,1,2,\ldots \tag{5.4}$$

Notice that for $\lambda = 1$, the confluent hypergeometric series (5.3) becomes the exponential series, i.e. we have ${}_1F_1[1;1;\theta] = \sum_{x=0}^{\infty}\theta^x/x! = e^\theta$, and distribution (5.4) reduces to the Poisson distribution with parameter θ. Also, $\lambda^{(0)} = 1$ for any $\lambda \in \mathbb{R}^+$. For fixed λ, the function $\lambda^{(x)}$ depends only on x, hence distribution (5.4) becomes a *power series distribution* with θ being a power parameter (cf. Noack, 1950).

In order to avoid difficulties by computing the confluent hypergeometric function ${}_1F_1[\cdot]$ it is helpful to use its relation to the *lower incomplete gamma function* $\gamma(a,x)$, formulated generally as (cf. Johnson et al., 1992, p. 14)

$$\gamma(a,x) = a^{-1}x^a e^{-x}{}_1F_1[1;a+1;x]\,, \tag{5.5}$$

where $a \neq 0,-1,-2,\ldots$ and $x > 0$. When $a = \lambda - 1$ and $x = \theta$ is substituted above, we obtain

$${}_1F_1[1;\lambda;\theta] = (\lambda-1)\theta^{1-\lambda}e^\theta\gamma(\lambda-1,\theta)\,, \tag{5.6}$$

provided that $\lambda \neq 1$ and $\theta > 0$ holds. Moreover, the function $\gamma(a,x)$ is closely related to the cumulative distribution function of a Gamma distribution, being specified by

$$F(x,a) = \frac{1}{\Gamma(a)}\int_0^x t^{a-1}e^{-t}dt = \frac{\gamma(a,x)}{\Gamma(a)}\,, \quad a > 0,\ x \geq 0\,. \tag{5.7}$$

Therefore, for all positive real values of θ and $\lambda > 1$, the function $\gamma(\lambda-1,\theta)$ can be computed by this relation as $\gamma(\lambda-1,\theta) = F(\theta,\lambda-1)\Gamma(\lambda-1)$. However, whenever $0 < \lambda < 1$, $F(\theta,\lambda-1)$ is not defined. In that case, for the calculation of $\gamma(\lambda-1,\theta)$ one should use its relationship to the *upper incomplete gamma function* $\Gamma(\lambda-1,\theta)$, expressed by $\gamma(\lambda-1,\theta) = \Gamma(\lambda-1) - \Gamma(\lambda-1,\theta)$. Nevertheless, there is a function `hyperg_1F1(a, b, x)` available in `R` package `gsl` that offers solutions for all real values of a, b and x, hence combines both cases above (see Hankin, 2006).

Crow and Bardwell (1965) stated that the parameter λ determines the type of the distribution (5.4), as it measures dispersion, and classified this model according to the values of λ. For $\lambda > 1$, they called it *super-Poisson* since the variance exceeds the mean, whereas if $0 < \lambda < 1$ the mean is greater than the variance, hence it has been named *sub-Poisson*. Obviously, when $\lambda = 1$ we have the equality of mean and variance, i.e. the standard Poisson case. The next two sections attest that the same holds for both the 1-displaced and the size-biased version of the HP distribution.

5.2 1-Displaced Hyper-Poisson Distribution

Let X^d be a discrete random variable that is 1-displaced HP distributed. Its pmf is then defined by the following formula (cf. Wimmer and Altmann, 1999, p. 281)

$$\pi^d_{x|\lambda,\theta} = P(X^d = x) = \frac{\theta^{x-1}}{{}_1F_1[1;\lambda;\theta]\lambda^{(x-1)}}\,,\quad x = 1, 2, \dots\,, \tag{5.8}$$

where $\lambda > 0$, $\theta > 0$ and ${}_1F_1[1;\lambda;\theta]$ is given by (5.3). Notice that for $\lambda = 1$ it follows that $\lambda^{(x-1)} = (x-1)!$ and ${}_1F_1[1;\lambda;\theta] = e^\theta$, hence as a result the 1-displaced Poisson model is obtained.

The pgf of distribution (5.8) can be derived directly by definition as follows

$$G_{X^d}(t) = \sum_{i=1}^{\infty} t^i \pi^d_{i|\lambda,\theta} = \frac{t}{{}_1F_1[1;\lambda;\theta]} \sum_{i=1}^{\infty} \frac{(\theta t)^{i-1}}{\lambda^{(i-1)}} = t\,\frac{{}_1F_1[1;\lambda;\theta t]}{{}_1F_1[1;\lambda;\theta]}\,. \tag{5.9}$$

Based on the partial derivatives of the confluent hypergeometric functions used above and given by the following relations

$$\begin{aligned} \frac{\partial^n {}_1F_1[1;\lambda;\theta]}{\partial\theta^n} &= \frac{n!}{\lambda^{(n)}}{}_1F_1[1+n;\lambda+n;\theta]\,, \\ \frac{\partial^n {}_1F_1[1;\lambda;\theta t]}{\partial t^n} &= \frac{n!\theta^n}{\lambda^{(n)}}{}_1F_1[1+n;\lambda+n;\theta t]\,, \end{aligned} \tag{5.10}$$

the mean of distribution (5.8), derived from the pgf above, results in

$$\mathrm{E}(X^d) = G'_{X^d}(1) = 1 + \frac{\theta}{\lambda}\frac{{}_1F_1[2;\lambda+1;\theta]}{{}_1F_1[1;\lambda;\theta]} = 1 + \frac{\theta}{{}_1F_1[1;\lambda;\theta]}\frac{\partial {}_1F_1[1;\lambda;\theta]}{\partial\theta}\,. \tag{5.11}$$

However, instead of the term ${}_1F_1[2;\lambda+1;\theta]$ we can use its relation to the function ${}_1F_1[1;\lambda;\theta]$ generally formulated as (cf. Abramowitz and Stegun, 1965, p. 506)

$$a {}_1F_1[a+1;b;z] = (1+a-b)\ {}_1F_1[a;b;z] + (b-1)\ {}_1F_1[a;b-1;z]\,, \tag{5.12}$$

which by substituting $a = 1$, $b = \lambda + 1$ and $z = \theta$ results in

$${}_1F_1[2;\lambda+1;\theta] = (1-\lambda)\,{}_1F_1[1;\lambda+1;\theta] + \lambda\,{}_1F_1[1;\lambda;\theta]\,, \tag{5.13}$$

where further ${}_1F_1[1;\lambda+1;\theta] = \lambda\theta^{-1}\left({}_1F_1[1;\lambda;\theta] - 1\right)$ holds. Consequently, we obtain the mean in the simplified form as follows[1]

$$\mu^d = \mathrm{E}(X^d) = 1 + \theta + (1-\lambda)(1 - {}_1F_1^{-1}[1;\lambda;\theta])\,. \tag{5.14}$$

[1] Evidently, as $\mathrm{E}(X^d) = \mathrm{E}(X) + 1$, we have $\mu = \mathrm{E}(X) = \theta + (1-\lambda)(1 - {}_1F_1^{-1}[1;\lambda;\theta])$ being the expression obtained by Bardwell and Crow (1964) for the mean of the HP distribution (5.2). However, as noted by the authors, they derived it differently by summing both sides of the recurrence relation for the successive HP probabilities over all x.

To derive the variance of distribution (5.8), we set $t = 1$ in the second derivative of $G_{X^d}(t)$ to first determine the second factorial moment. Using the second relation of (5.10) and the fact that $\partial/\partial t({}_1F_1[2;\lambda+1;\theta t]\,) = 2\theta{}_1F_1[3;\lambda+2;\theta t]/(\lambda+1)$ we get

$$\mu^d_{(2)} = G''_{X^d}(1) = \frac{2\theta}{\lambda}\frac{{}_1F_1[2;\lambda+1;\theta]}{{}_1F_1[1;\lambda;\theta]} + \frac{2\theta^2}{\lambda(\lambda+1)}\frac{{}_1F_1[3;\lambda+2;\theta]}{{}_1F_1[1;\lambda;\theta]}\,. \tag{5.15}$$

Replacing subsequently ${}_1F_1[3;\lambda+2;\theta]$ and ${}_1F_1[2;\lambda+2;\theta]$ by the expressions resulting from relation (5.12) when $(a,b,z) = (2,\lambda+2,\theta)$ and $(a,b,z) = (1,\lambda+2,\theta)$, respectively, and ${}_1F_1[2;\lambda+1;\theta]$ by equation (5.13) yields after few simplifications

$$\mu^d_{(2)} = (\mu^d - 1)(2\theta + 2 - \lambda) + 2\theta - \theta\mu^d\,. \tag{5.16}$$

Finally, the variance is obtained as

$$\operatorname{var}(X^d) = \mu^d_{(2)} - \mu^d(\mu^d - 1) = \theta\mu^d + (\mu^d - 1)(2 - \mu^d - \lambda)\,. \tag{5.17}$$

Calculation of the higher factorial moments becomes quite tedious and time consuming using the method described above. Hence, we rather suggest to combine the recurrence formula (2.38), Section 2.3.1 with formula (B.8), Appendix B, to get

$$\mu^d_{(k)} = \sum_{i=0}^{k} s(k,i)\mu'_i + k\sum_{i=0}^{k-1} s(k-1,i)\mu'_i\,, \tag{5.18}$$

where $s(k,i)$ are the Stirling numbers of the first kind[2] and μ'_i denotes the i-th raw moment of distribution (5.4), given recursively by (cf. Crow and Bardwell, 1965)

$$\mu'_{k+1} = (\theta - \lambda + 1)\mu'_k + \theta\sum_{i=1}^{k}\binom{k}{i}\mu'_{k-i}\,, \quad \text{for } k = 1,2,\ldots \tag{5.19}$$

All moments of distribution (5.8) for $\lambda,\theta \in \mathbb{R}^+$ exist. With factorial moments determined, raw and central moments of distribution (5.8) can be calculated using the relations (B.9) and (B.13) from Appendix B. In particular, the second and third raw moments are derived from (5.18) as follows

$$\begin{aligned} \mu^{d'}_2 &= \mu^d_{(2)} + \mu^d = (\theta - \lambda + 3)\mu^d + \lambda - 2\,,\\ \mu^{d'}_3 &= \mu^d_{(3)} + 3\mu^d_{(2)} + \mu^d = (\theta - \lambda + 4)\mu'^d_2 + (2\lambda - 5)\mu^d - \lambda + 2\,. \end{aligned} \tag{5.20}$$

The basic properties of the distribution above in relation to those of its original and size-biased (see forthcoming section) versions are summarized in Table 5.2.

To study further the behavior of distribution (5.8), the successive probabilities for possible values of x by varying values of parameters λ and θ are computed using the following recurrence relation

$$\pi^d_{x|\lambda,\theta} = \frac{\theta}{\lambda + x - 2}\,\pi^d_{x-1|\lambda,\theta}\,, \quad \text{for } x \geq 2\,, \tag{5.21}$$

[2] For more details on Stirling numbers of the first kind see Appendix B.

where $\pi^d_{1|\lambda,\theta} = {}_1F_1^{-1}[1;\lambda;\theta]$, as defined by formula (5.8). Figure 5.1 contains six plots for $\theta = 0.8$ (left column), $\theta = 2.4$ (right column), and $\lambda = 0.3, 1, 4.3$ and illustrates changes in the form of distribution (5.8) when for any given value of λ the value of θ changes or when θ is fixed but parameter λ varies. The comparison to

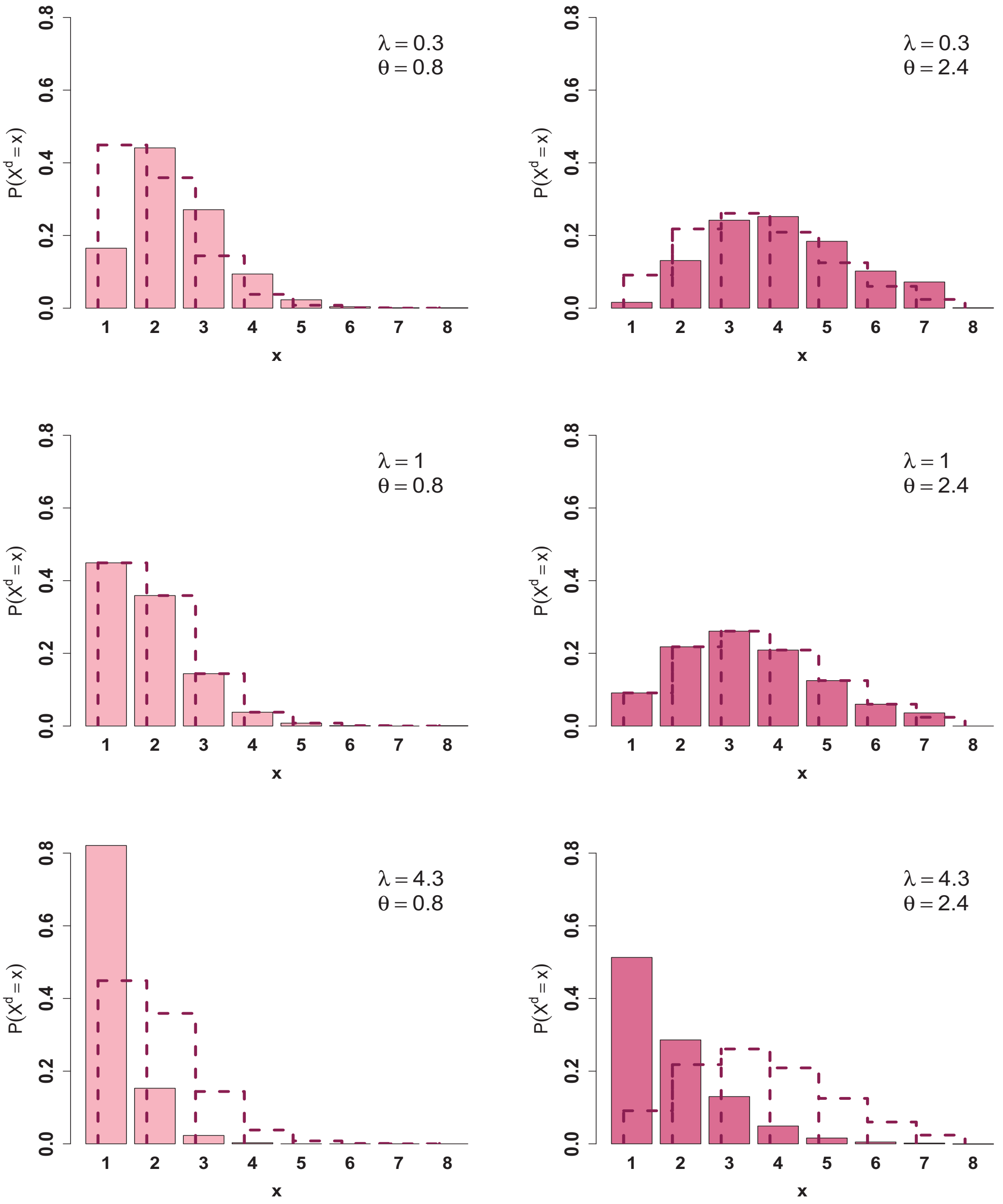


Figure 5.1: Graphs of the HP(λ, θ) and the Poisson(θ) (dashed bars) probability distributions for $\theta = 0.8$ (left column), $\theta = 2.4$ (right column) and different values of λ.

the 1-displaced Poisson distribution with parameter θ, displayed here as dashed bar plot, is given to provide an insight into the boundary case. Evidently, as λ gets close to zero, the hypergeometric series (5.3) becomes positively infinite, and hence the probability $\pi_{1|\lambda,\theta}$ approaches zero. When θ is fixed, the series representation of the function ${}_1F_1[1;\lambda;\theta]$ decreases as λ increases, and becomes unity for λ large enough. As a result, we have $\pi_{1|\lambda,\theta} \to 1$, therefore the limiting form of distribution (5.8) when $\lambda \to \infty$ is a degenerate distribution with all its mass concentrated at one. From the last row in Figure 5.1 it is clear that it happens faster for smaller values of θ. Notice also that the asymmetry in the probability distribution decreases as the value of θ increases. For $\lambda = 1$ the 1-displaced HP model achieves the Poisson limit, as obvious from the middle row in Figure 5.1, whereas for both $0 < \lambda < 1$ and $\lambda > 1$ we have departure from the Poisson case.

To summarize, parameter λ carries information about the type of the distribution above. In order to verify it, we calculate the index of dispersion

$$\delta = \frac{\mathrm{var}(X^d)}{\mathrm{E}(X^d) - 1} = \theta - \lambda - \mu^d + 2 + \frac{\theta}{\mu^d - 1}\,, \tag{5.22}$$

where μ^d is given by (5.14). Table 5.1 shows values of μ^d and δ for various combinations of the two parameters. For $\lambda = 1$ we have $\delta = 1$, since then $\mathrm{E}(X^d) = 1 + \theta$ and $\mathrm{var}(X^d) = \theta$. Clearly, when $0 < \lambda < 1$, δ is strictly below 1, whereas it reaches maximum at unity for large enough θ. Therefore, we have underdispersion with respect to Poisson variation. However, when $\lambda > 1$ the distribution becomes overdispersed.

Table 5.1: Under- and overdispersion in the 1-displaced Hyper-Poisson distribution

		θ						
λ		0.1	0.5	0.8	1	2.4	5	8
	μ^d	1.29	2.00	2.38	2.62	4.09	6.70	9.70
0.3	δ	0.87	0.70	0.69	0.70	0.79	0.88	0.92
	μ^d	1.16	1.71	2.08	2.31	3.78	6.40	9.40
0.6	δ	0.96	0.89	0.87	0.86	0.88	0.93	0.95
	μ^d	1.11	1.54	1.86	2.07	3.49	6.10	9.10
0.9	δ	0.99	0.98	0.97	0.97	0.97	0.98	0.99
	μ^d	1.10	1.50	1.80	2.00	3.40	6.00	9.00
1	δ	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	μ^d	1.08	1.40	1.66	1.83	3.14	5.70	8.70
1.3	δ	1.01	1.04	1.06	1.06	1.08	1.06	1.04
	μ^d	1.02	1.13	1.21	1.27	1.79	3.31	5.82
4.3	δ	1.01	1.07	1.12	1.15	1.34	1.56	1.54

Notice also that as λ increases, the mean decreases, hence the value of δ increases as well, irrespective of the value of θ. Still, when λ becomes quite large the distribution (5.8) transforms to a degenerate distribution concentrated at one (cf. Figure 5.1 above), hence its variance becomes zero, whereas the mean becomes unity. For this reason δ is then indefinite.

5.3 Size-Biased Hyper-Poisson Distribution

The size-biased HP distribution results by applying transformation (2.41), Section 2.3.2 to the original HP distribution (5.4), where $\mathrm{E}(X)$ is obtained from (5.11). The pmf of the corresponding random variable X^* is then for $x = 1, 2, \ldots$ given by

$$\pi^*_{x|\lambda,\theta} = P(X^* = x) = \frac{x\theta^x}{{}_1F_1[1;\lambda;\theta]\lambda^{(x)}} \frac{\lambda\,{}_1F_1[1;\lambda;\theta]}{\theta\,{}_1F_1[2;\lambda+1;\theta]} = \frac{x\lambda\theta^{x-1}}{{}_1F_1[2;\lambda+1;\theta]\lambda^{(x)}} \,. \tag{5.23}$$

Although its pgf might be derived directly from (5.23), another possibility is to use relation (2.42), Section 2.3.2, where $G'_X(t)$ is obtained applying the second equation of (5.10). As a result, we have

$$G_{X^*}(t) = t\,\frac{{}_1F_1[2;\lambda+1;\theta t]}{{}_1F_1[2;\lambda+1;\theta]} \,. \tag{5.24}$$

All raw moments of the HP distribution (5.23) for $\lambda, \theta \in \mathbb{R}^+$ exist, and can easily be computed from those of the original HP distribution specified by (5.19), by utilizing relation (2.46), Section 2.3.2 where $\mathrm{E}(X)$ is determined from (5.11) as $\mathrm{E}(X^d) - 1$. Accordingly, the mean is found to be

$$\mu^* = \mathrm{E}(X^*) = (\theta - \lambda + 1) + \lambda \frac{{}_1F_1[1;\lambda;\theta]}{{}_1F_1[2;\lambda+1;\theta]} \,. \tag{5.25}$$

When this relation is rewritten as $\lambda_1 F_1[1;\lambda;\theta]/{}_1F_1[2;\lambda+1;\theta] = \mu^* - (\theta - \lambda + 1)$ and substituted into (2.46) the second and third raw moments are obtained as

$$\begin{aligned} \mu_2^{*\prime} &= (\theta + \lambda - 1) + (\theta - \lambda + 2)\mu^* \,, \\ \mu_3^{*\prime} &= (\theta - \lambda + 1)\mu_2^{*\prime} + (3\theta + 1)\mu^* + 2\theta + \lambda - 1 \,. \end{aligned} \tag{5.26}$$

Consequently, after few algebraic simplifications the variance results in

$$\mathrm{var}(X^*) = \mu_2^{*\prime} - (\mu^*)^2 = \theta(\mu^* + 1) + (\mu^* - 1)(1 - \mu^* - \lambda) \,. \tag{5.27}$$

With the size-biased raw moments calculated, the factorial and central moments can be determined using relations (B.8) and (B.13) from Appendix B, respectively.

Figure 5.2 illustrates the differences in the size-biased HP probability distribution compared to the corresponding 1-displaced version for $\theta = 0.8, 2.4$ and $\lambda = 0.1, 1, 5.3$. In order to draw individual size-biased probabilities we use the ratio of the recursive probabilities given by

$$\pi^*_{x|\lambda,\theta} = \frac{x\theta}{(x-1)(\lambda + x - 1)} \pi^*_{x-1|\lambda,\theta} \,, \quad \text{for } x \geq 2 \,, \tag{5.28}$$

where $\pi^*_{1|\lambda,\theta} = {}_1F_1^{-1}[2;\lambda+1;\theta]$, as defined by (5.23). It is evident from the graphs that as long as $0 < \lambda < 1$ we have $P(X^d = 1) < P(X^* = 1)$ and both probabilities

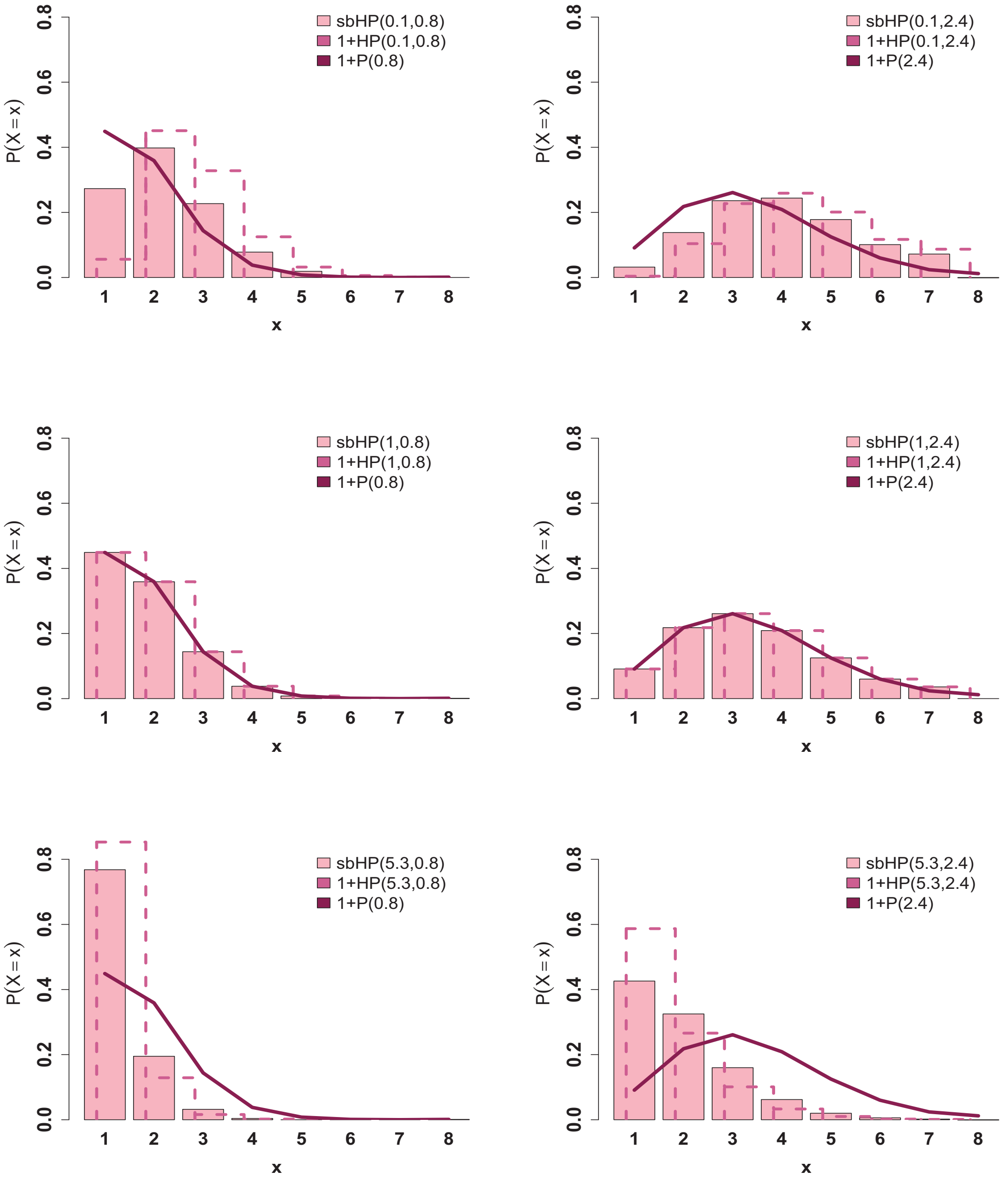


Figure 5.2: The size-biased HP (sbHP(λ,θ)) and 1-displaced HP (1+HP(λ,θ)) distributions compared to the 1-displaced Poisson (1+P(θ)) for $\theta = 0.8, 2.4$ and $\lambda = 0.1, 1, 5.3$.

are smaller than $e^{-\theta}$, hence there is *one-deflation* compared to the Poisson model. Notice that this also affects the probabilities of the remaining frequencies. All three distributions coincide for $\lambda = 1$. For $\lambda > 1$ we have $P(X^d = 1) > P(X^* = 1) > e^{-\theta}$, i.e. *one-inflation* with respect to the Poisson case. It may also be noted that both HP distributions exhibit opposite behavior to that observed for the SP distribution.

Table 5.2: Hyper-Poisson distribution: original, 1-displaced and size-biased forms

	Random Variable		
	X	X^d	X^*
Notation	$\mathrm{HP}(\lambda,\theta)$	$1{+}\mathrm{HP}(\lambda,\theta)$	$\mathrm{sbHP}(\lambda,\theta)$
Range	$\mathbb{N}_0$	$\mathbb{N}$	$\mathbb{N}$
pmf	$\dfrac{\theta^x}{{}_1F_1[1;\lambda;\theta]\lambda^{(x)}}$	$\pi_{x-1\vert\lambda,\theta}$	$\dfrac{x\pi_{x\vert\lambda,\theta}}{\mathrm{E}(X)}$
pgf	$\dfrac{{}_1F_1[1;\lambda;\theta t]}{{}_1F_1[1;\lambda;\theta]}$	$tG_X(t)$	$t\,\dfrac{{}_1F_1[2;\lambda+1;\theta t]}{{}_1F_1[2;\lambda+1;\theta]}$
$\mathrm{E}(\cdot)$	$\theta+(1-\lambda)(1-{}_1F_1^{-1}[1;\lambda;\theta])$	$1+\mathrm{E}(X)$	$(\theta-\lambda+1)+\lambda\dfrac{{}_1F_1[1;\lambda;\theta]}{{}_1F_1[2;\lambda+1;\theta]}$
$\mathrm{var}(\cdot)$	$\theta(1+\mu)+\mu(1-\mu-\lambda)$	$\mathrm{var}(X)$	$\theta(1+\mu^*)+(\mu^*-1)(1-\mu^*-\lambda)$

Based on the graphs above we can conclude that the parameter λ defines the type of distribution (5.23). To measure effective dispersion we apply the relations (5.25) and (5.27) and get the index of dispersion by the following formula

$$\delta = \frac{\mathrm{var}(X^*)}{\mathrm{E}(X^*)-1} = \theta-\lambda-\mu^*+1+\frac{2\theta}{\mu^*-1}\,. \tag{5.29}$$

Since the ability of this distribution to model under-, equi- and overdispersed sample data is not apparent from the expression above, we compute δ and μ^* for different values of λ and θ. Table 5.3 displays the results obtained. It can be observed that

Table 5.3: Under- and overdispersion in the size-biased Hyper-Poisson distribution

		θ						
λ		0.1	0.5	0.8	1	2.4	5	8
0.3	μ^*	1.15	1.70	2.08	2.32	3.88	6.58	9.62
	δ	0.98	0.93	0.91	0.90	0.89	0.92	0.94
0.6	μ^*	1.12	1.60	1.94	2.17	3.66	6.33	9.35
	δ	0.99	0.97	0.95	0.95	0.94	0.95	0.96
0.9	μ^*	1.11	1.52	1.83	2.04	3.46	6.08	9.09
	δ	1.00	0.99	0.99	0.99	0.99	0.99	0.99
1	μ^*	1.10	1.50	1.80	2.00	3.40	6.00	9.00
	δ	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.3	μ^*	1.09	1.44	1.71	1.90	3.22	5.76	8.74
	δ	1.00	1.02	1.02	1.03	1.04	1.04	1.03
4.3	μ^*	1.04	1.20	1.33	1.42	2.13	3.87	6.36
	δ	1.01	1.05	1.08	1.09	1.21	1.32	1.33

analogous to the 1-displaced HP model and depending on the fact whether $0 < \lambda < 1$ or $\lambda > 1$, we have under- or overdispersion, respectively. Obviously, the 1-displaced Poisson limit is obtained for $\lambda = 1$.

5.4 Parameter Estimation

This section deals with parameter estimation of the two different versions of the HP distribution, defined by (5.8) and (5.23). Since their estimators are different we discuss the three estimation methods, mentioned in Section 2.4, for each of them separately. Furthermore, alternatives to these methods are also considered here.

5.4.1 Estimation by Method of Moments

1-displaced HP model

(a) λ **known**: The moment estimator of parameter θ is obtained most easily from equation (5.17) by replacing the first two theoretical moments by the corresponding sample moments. Consequently, we obtain

$$\hat{\theta}_{\mathrm{MM}} = \frac{m_2 - (\bar{x}-1)(2-\bar{x}-\lambda)}{\bar{x}} \,. \tag{5.30}$$

(b) λ **unknown**: In this situation the estimation procedure becomes more complex. The moment estimators of λ and θ follow then by equating the sample mean $\bar{x}$ and the second raw moment m_2' to the corresponding theoretical moments, specified by (5.14) and the first equation of (5.20), respectively. Thus, we get the following system of equations

$$\bar{x} = 1 + \theta + (1-\lambda)(1 - {}_1F_1^{-1}[1;\lambda;\theta]) \,, \;\; m_2' = (\theta - \lambda + 3)\bar{x} + \lambda - 2 \,. \tag{5.31}$$

Eliminating parameter λ from the previous equations, we obtain as a result

$$H(\theta) = {}_1F_1\left[1; \frac{\theta\bar{x} + 3\bar{x} - m_2' - 2}{\bar{x}-1}; \theta\right] - \frac{\theta\bar{x} + 2\bar{x} - m_2' - 1}{\theta + \bar{x}^2 - m_2'} = 0 \,. \tag{5.32}$$

This equation may be solved iteratively for θ using, e.g. function `uniroot()` available in `R` software. The positive root of equation (5.32) is the required estimate. With $\hat{\theta}_{\mathrm{MM}}$ calculated in this way, $\hat{\lambda}_{\mathrm{MM}}$ follows from the second equation of (5.31) as

$$\hat{\lambda}_{\mathrm{MM}} = \frac{\hat{\theta}_{\mathrm{MM}}\,\bar{x} + 3\bar{x} - m_2' - 2}{\bar{x}-1} \,. \tag{5.33}$$

However, simple moment estimators are obtained when the first equation of the system (5.31) is replaced by that for the third raw moment defined by (5.20). Solving simultaneously thus determined system of equations, the estimator for θ follows as

$$\hat{\theta}_{\mathrm{MM}} = \frac{m_3'(\bar{x}-1) + m_2'\bar{x} + m_2' - (m_2')^2 - \bar{x}^2}{2\bar{x}^2 - \bar{x} - m_2'} \,, \tag{5.34}$$

whereas for such $\hat{\theta}_{\mathrm{MM}}$, the moment estimator of λ is again given by (5.33).

size-biased HP model

(a) λ **known**: The moment estimator of θ can be found from equation (5.27) as

$$\hat{\theta}_{\text{MM}} = \frac{m_2 - (\bar{x}-1)(1-\bar{x}-\lambda)}{\bar{x}+1}\,. \tag{5.35}$$

(b) λ **unknown**: Contrary to (a), the estimation procedure here is more complex. Starting with the first equation of (5.26), and substituting the theoretical moments by the corresponding sample moments we obtain the following expression for λ

$$\lambda = \frac{\theta(\bar{x}+1) + 2\bar{x} - m_2' - 1}{\bar{x}-1}\,. \tag{5.36}$$

The second equation of the system is given by equating the population mean, defined by (5.25), to its empirical counterpart $\bar{x}$ which after some transformations results in

$$(\bar{x} - \theta + \lambda - 1)\,{}_1F_1[2;\lambda+1;\theta] - \lambda\,{}_1F_1[1;\lambda;\theta] = 0\,. \tag{5.37}$$

Substituting subsequently λ by expression (5.36) we get a quite complex equation in θ which can be solved iteratively in the similar manner as in the 1-displaced case. With $\hat{\theta}_{\text{MM}}$ calculated we ultimately obtain $\hat{\lambda}_{\text{MM}}$ from (5.36).

To derive simple estimators of λ and θ we return to equation (5.36) and combine it with the second equation of (5.26) instead of that for the population mean. After further simplifications, the estimator of θ is expressed by

$$\hat{\theta}_{\text{MM}} = \frac{m_3'(\bar{x}-1) - (m_2')^2 + \bar{x}m_2' - \bar{x}^2 + m_2'}{3\bar{x}^2 - 2m_2' - 1}\,, \tag{5.38}$$

whereas $\hat{\lambda}_{\text{MM}}$ results from (5.36) when θ is replaced with $\hat{\theta}_{\text{MM}}$.

5.4.2 Estimation by Maximum Likelihood

1-displaced HP model

Consider a random sample of size n from the 1-displaced HP model, defined by (5.8). To find the maximum likelihood (ML) estimator of the parameter vector $\mathbf{\Theta} = (\lambda, \theta)$ we begin with the following likelihood function

$$L(\lambda, \theta | f_1, \ldots, f_k) = \prod_{i=1}^{k} \left(\frac{\theta^{i-1}\Gamma(\lambda)}{{}_1F_1[1;\lambda;\theta]\Gamma(\lambda+i-1)} \right)^{f_i}, \tag{5.39}$$

where f_i are the observed frequencies such that $\sum_{i=1}^{k} f_i = n$ with k being the largest observation and maximize its log-likelihood function given by

$$\begin{aligned} l(\lambda, \theta | f_1, \ldots, f_k) &= \log L(\lambda, \theta | f_1, \ldots, f_k) = \\ &= n(\bar{x}-1)\log\theta + n\log\Gamma(\lambda) - n\log{}_1F_1[1;\lambda;\theta] - \sum_{i=1}^{k} f_i \log\Gamma(\lambda+i-1)\,. \end{aligned} \tag{5.40}$$

Then, we equate the first partial derivatives of (5.40) with respect to the parameters λ and θ to zero, which by substituting the *digamma function* $\Psi(x) = \Gamma'(x)/\Gamma(x)$ enables us to obtain the following score equations

$$\begin{aligned} \frac{\partial l(\lambda,\theta|\mathbf{f})}{\partial\lambda} &= n\Psi(\lambda) - \frac{n}{{}_1F_1[1;\lambda;\theta]}\frac{\partial_1 F_1[1;\lambda;\theta]}{\partial\lambda} - \sum_{i=1}^{k} f_i\Psi(\lambda+i-1) = 0\,, \\ \frac{\partial l(\lambda,\theta|\mathbf{f})}{\partial\theta} &= \frac{n(\bar{x}-1)}{\theta} - \frac{n}{{}_1F_1[1;\lambda;\theta]}\frac{\partial_1 F_1[1;\lambda;\theta]}{\partial\theta} = 0\,, \text{ with } \mathbf{f} = (f_1,\dots,f_k)\,. \end{aligned} \tag{5.41}$$

(a) λ known: The first equation of the system above becomes in this case irrelevant, whereas the second one, by using the first relation of (5.10), simplifies to

$$\lambda(\bar{x}-1)_1F_1[1;\lambda;\theta] - \theta_1F_1[2;\lambda+1;\theta] = 0\,. \tag{5.42}$$

Hence, the estimator $\hat{\theta}_{\text{ML}}$ is obtained by solving (5.42) iteratively for θ.

(b) λ unknown: The system (5.41) may be solved iteratively for the unknown parameters using the Newton-Raphson algorithm. However, from the definition of the function ${}_1F_1[1;\lambda;\theta]$ given in (5.3) it is apparent that the calculation of its partial derivatives with respect to λ is quite troublesome. Apart from the series representation, the confluent hypergeometric function ${}_1F_1[1;\lambda;\theta]$ can be defined using the following integral (cf. Johnson et al., 1992, p. 20)

$${}_1F_1[1;\lambda;\theta] = \frac{\Gamma(\lambda)}{\Gamma(\lambda-1)}\int_0^1 (1-p)^{\lambda-2}e^{\theta p}dp\,, \tag{5.43}$$

where $\Gamma(\cdot)$ is the gamma function. Both representations of ${}_1F_1[1;\lambda;\theta]$ are inappropriate for the analytical treatment needed in maximum likelihood estimation. The contribution of Butler and Wood (2002) offers a solution to this problem. The authors derived a Laplace approximation for the integral in (5.43), both for the scalar and matrix case, and based on simulations verified extremely high precision of their approach. Generally, they considered the integral

$$I = \int_{p\in D} h(p)e^{-\omega g(p)}dp, \tag{5.44}$$

where $D \subseteq R^d$ and ω is a real parameter. If $g(p)$ has a unique minimum at the stationary point $p_s \in D$ of $g(p)$, then Laplace's approximation of I is given by

$$\tilde{I} = (2\pi)^{d/2}\omega^{-d/2}|g''(p_s)|^{-1/2}h(p_s)e^{-\omega g(p_s)}\,. \tag{5.45}$$

For our purposes, only the scalar case is of interest and we take $\omega = 1$ for simplicity. Moreover, the authors remarked that when implementing Laplace's approximation, the choice of the functions g and h as well as the possibility of calibration (see below) can affect its exactness. To approximate the integral in (5.43) they proposed to use the following representation of the functions g and h

$$\begin{aligned} h(p) &= B(1,\lambda-1)^{-1}p^{-1}(1-p)^{-1}\,,\\ g(p) &= -\Big(\log p + (\lambda-1)\log(1-p) + \theta p\Big)\,, \end{aligned} \tag{5.46}$$

where $D = (0,1)$, $d = 1$ and $B(1,\lambda-1) = \Gamma(\lambda-1)/\Gamma(\lambda)$ denotes the beta function. Applying (5.45) the *raw Laplace approximation* of the function ${}_1F_1[1;\lambda;\theta]$ results in

$${}_1\widetilde{F}_1[1;\lambda;\theta] = \frac{(2\pi)^{1/2}p_s(1-p_s)^{\lambda-1}e^{\theta p_s}}{B(1,\lambda-1)\sqrt{(1-p_s)^2+(\lambda-1)p_s^2}}\,, \tag{5.47}$$

where the stationary point $p_s \in (0,1)$ obtained from solving $g'(p) = 0$ is given by

$$p_s = \frac{2}{\lambda-\theta+\sqrt{(\theta-\lambda)^2+4\theta}}\,. \tag{5.48}$$

Butler and Wood (2002) also noticed that it would be even better to use the approximation ${}_1\widehat{F}_1[1;\lambda;\theta]$ which is calibrated at $\theta = 0$, since it gives more accurate results. The *calibrated approximation* is defined by

$${}_1\widehat{F}_1[1;\lambda;\theta] = \frac{{}_1\widetilde{F}_1[1;\lambda;\theta]}{{}_1\widetilde{F}_1[1;\lambda;0]}\,, \tag{5.49}$$

where ${}_1\widetilde{F}_1[1;\lambda;0]$ has the same form as ${}_1\widetilde{F}_1[1;\lambda;\theta]$, but the stationary point $p_0 = 1/\lambda$ instead of p_s. Notice that p_0 is given by substituting $\theta = 0$ in expression (5.48). Eventually, the calibrated approximation in (5.49) can be written as

$${}_1\widehat{F}_1[1;\lambda;\theta] = \frac{p_s(1-p_s)^{\lambda-1}\sqrt{(1-p_0)^2+(\lambda-1)p_0^2}}{p_0(1-p_0)^{\lambda-1}\sqrt{(1-p_s)^2+(\lambda-1)p_s^2}}\,e^{\theta p_s} = \frac{A_{\lambda,\theta}}{B_{\lambda,\theta}}\,e^{\theta p_s}\,. \tag{5.50}$$

Hence, estimation and inference of the HP model (5.8) is based on the approximate log-likelihood that results from replacing ${}_1F_1[1;\lambda;\theta]$ in (5.40) by ${}_1\widehat{F}_1[1;\lambda;\theta]$. Consequently, the first partial derivatives are obtained when instead of $\partial\log{}_1F_1[1;\lambda;\theta]/\partial\lambda$ and $\partial\log{}_1F_1[1;\lambda;\theta]/\partial\theta$ in (5.41) one uses the expressions

$$\begin{aligned} \frac{\partial\log{}_1F_1[1;\lambda;\theta]}{\partial\lambda} &= \frac{\partial A_{\lambda,\theta}}{\partial\lambda}\frac{1}{A_{\lambda,\theta}} - \frac{\partial B_{\lambda,\theta}}{\partial\lambda}\frac{1}{B_{\lambda,\theta}} + \theta\,\frac{\partial p_s}{\partial\lambda}\,,\\ \frac{\partial\log{}_1F_1[1;\lambda;\theta]}{\partial\theta} &= \frac{\partial A_{\lambda,\theta}}{\partial\theta}\frac{1}{A_{\lambda,\theta}} - \frac{\partial B_{\lambda,\theta}}{\partial\theta}\frac{1}{B_{\lambda,\theta}} + p_s + \theta\,\frac{\partial p_s}{\partial\theta}\,. \end{aligned} \tag{5.51}$$

Subsequently, the second partial derivatives of the approximate log-likelihood needed for the Newton-Raphson algorithm are obtained as

$$\begin{aligned}
\frac{\partial^2 \widehat{l}(\lambda,\theta|f_1,\ldots,f_k)}{\partial\lambda^2} &= n\Psi'(\lambda) - n\frac{\partial^2 \log {}_1\widehat{F}_1[1;\lambda;\theta]}{\partial\lambda^2} - \sum_{i=1}^{k} f_i \Psi'(\lambda+i-1)\,, \\
\frac{\partial^2 \widehat{l}(\lambda,\theta|f_1,\ldots,f_k)}{\partial\lambda\partial\theta} &= -n\frac{\partial^2 \log {}_1\widehat{F}_1[1;\lambda;\theta]}{\partial\lambda\partial\theta}\,, \\
\frac{\partial^2 \widehat{l}(\lambda,\theta|f_1,\ldots,f_k)}{\partial\theta^2} &= -\frac{n(\bar{x}-1)}{\theta^2} - n\frac{\partial^2 \log {}_1\widehat{F}_1[1;\lambda;\theta]}{\partial\theta^2}\,,
\end{aligned} \tag{5.52}$$

where $\Psi'(x) = \partial\Psi(x)/\partial x$ denotes the *trigamma function*. The second derivatives of $\log {}_1\widehat{F}_1[1;\lambda;\theta]$ result from (5.51). Further details are omitted here, as the definitions of $A_{\lambda,\theta}$, $B_{\lambda,\theta}$ and p_s are extremely complex to write the explicit versions of their first and second derivatives.

However, the ML estimate $\widehat{\mathbf{\Theta}}_{\text{ML}}$ can also be found directly from the log-likelihood function specified in (5.40) applying e.g. the `R` function `optim()` for numerical optimization. In that case we do not need to find analytically the score function or the Hessian, since they are provided numerically by the procedure. Inference is then based on the estimated variance-covariance matrix derived from the Hessian.

size-biased HP model

The likelihood function of the size-biased HP model (5.23) is given by

$$L(\lambda,\theta|f_1,\ldots,f_k) = \prod_{i=1}^{k}\left(\frac{i\theta^{i-1}\Gamma(\lambda+1)}{{}_1F_1[2;\lambda+1;\theta]\Gamma(\lambda+i)}\right)^{f_i}. \tag{5.53}$$

To find ML estimates of λ and θ we take logarithms of (5.53), differentiate the log-likelihood function $l(\lambda,\theta|f_1,\ldots,f_k) = \log L(\lambda,\theta|f_1,\ldots,f_k)$ with respect to λ and θ, equate to zero and simplify to get the following score equations

$$\begin{aligned}
\frac{\partial l(\lambda,\theta|\mathbf{f})}{\partial\lambda} &= n\Psi(\lambda+1) - \frac{n}{{}_1F_1[2;\lambda+1;\theta]}\frac{\partial {}_1F_1[2;\lambda+1;\theta]}{\partial\lambda} - \sum_{i=1}^{k} f_i\Psi(\lambda+i) = 0\,, \\
\frac{\partial l(\lambda,\theta|\mathbf{f})}{\partial\theta} &= \frac{n(\bar{x}-1)}{\theta} - \frac{n}{{}_1F_1[2;\lambda+1;\theta]}\frac{\partial {}_1F_1[2;\lambda+1;\theta]}{\partial\theta} = 0\,,
\end{aligned} \tag{5.54}$$

where $\mathbf{f} = (f_1,\ldots,f_k)$.

(a) λ known: The first equation of the system above becomes here meaningless. We obtain $\hat{\theta}_{\text{ML}}$ by solving iteratively for θ the following simplified form of the second equation above

$$\lambda(\bar{x}-1){}_1F_1[2;\lambda+1;\theta] - 2\theta {}_1F_1[3;\lambda+2;\theta] = 0\,. \tag{5.55}$$

(b) λ unknown: When there is no information about λ, to solve the system (5.54) for the unknown parameters iteratively we use findings of Butler and Wood (2002) to get derivatives of the function ${}_1F_1[2;\lambda+1;\theta]$. In a similar manner as in the 1-displaced case we can show that its calibrated approximation is given by

$$
{}_1\widehat{F}_1[2;\lambda+1;\theta] = \frac{p_s^2(1-p_s)^{\lambda-1}\sqrt{2(1-p_0)^2+(\lambda-1)p_0^2}}{{p_0}^2(1-p_0)^{\lambda-1}\sqrt{2(1-p_s)^2+(\lambda-1)p_s^2}}\, e^{\theta p_s} = \frac{C_{\lambda,\theta}}{D_{\lambda,\theta}}\, e^{\theta p_s}\,, \quad (5.56)
$$

where $p_s = 4/(\lambda-\theta+1+\sqrt{(\theta-\lambda-1)^2+8\theta}\,)$ and $p_0 = 2/(\lambda+1)$. Consequently, the estimation and inference are based on the approximated log-likelihood $\widehat{l}(\lambda,\theta|f_i)$ where ${}_1\widehat{F}_1[2;\lambda+1;\theta]$ is used instead of ${}_1F_1[2;\lambda+1;\theta]$. The score equations are now obtained by replacing the partial derivatives of $\log{}_1F_1[2;\lambda+1;\theta]$ with respect to λ and θ in (5.54) with the following relations

$$
\begin{aligned}
\frac{\partial \log{}_1\widehat{F}_1[2;\lambda+1;\theta]}{\partial\lambda} &= \frac{\partial C_{\lambda,\theta}}{\partial\lambda}\,\frac{1}{C_{\lambda,\theta}} - \frac{\partial D_{\lambda,\theta}}{\partial\lambda}\,\frac{1}{D_{\lambda,\theta}} + \theta\,\frac{\partial p_s}{\partial\lambda}\,,\\
\frac{\partial \log{}_1\widehat{F}_1[2;\lambda+1;\theta]}{\partial\theta} &= \frac{\partial C_{\lambda,\theta}}{\partial\theta}\,\frac{1}{C_{\lambda,\theta}} - \frac{\partial D_{\lambda,\theta}}{\partial\theta}\,\frac{1}{D_{\lambda,\theta}} + p_s + \theta\,\frac{\partial p_s}{\partial\theta}\,.
\end{aligned} \quad (5.57)
$$

The second partial derivatives of the approximate log-likelihood are derived applying the derivatives of (5.57). The details of their explicit versions are omitted because of their complexity.

Furthermore, the size-biased HP model (5.23) can be implemented in software R, using the functions for numerical optimization, as already mentioned for the 1-displaced HP case. Some of them allow even to choose whether numerical or rather analytical gradient and Hessian should be used.

5.4.3 Estimation Based on Mean and First Frequency Class

1-displaced HP model

By equating the sample mean $\bar{x}$ and the relative frequency f_1/n for the first class to the population mean μ^d and the theoretical probability $\pi^d_{1|\lambda,\theta}$, respectively, we get the following estimating equations for the 1-displaced HP model, defined by (5.8)

$$
\bar{x} = 1+\theta+(1-\lambda)(1-{}_1F_1^{-1}[1;\lambda;\theta])\,, \;\text{ and }\; f_1/n = {}_1F_1^{-1}[1;\lambda;\theta]\,. \quad (5.58)
$$

(a) λ known: In this particular case the estimator of θ can be determined from the first equation of (5.58) as follows

$$
\hat{\theta}_{\text{FF}} = \bar{x}-1-(1-\lambda)(1-f_1/n)\,. \quad (5.59)
$$

(b) λ unknown: The estimators $\hat{\lambda}_{\text{FF}}$ and $\hat{\theta}_{\text{FF}}$ can not be obtained here as the simultaneous solutions of the previous system. Rather, their solution requires additional estimating equation. Following the approach suggested by Bardwell and Crow (1964) we equate the second sample moment m_2 to the corresponding theoretical one, given by (5.17), to get

$$m_2 = \theta \bar{x} + (\bar{x} - 1)(2 - \bar{x} - \lambda)\,. \tag{5.60}$$

When ${}_1F_1^{-1}[1; \lambda; \theta]$ is eliminated from the two equations of (5.58), we then substitute the obtained expression for θ in (5.60). After some algebraic simplifications the explicit parameter estimators of the 1-displaced HP model (5.8) are

$$\hat{\lambda}_{\text{FF}} = \frac{n(m_2 + 2) - \bar{x}(n + f_1)}{n - \bar{x} f_1}\,, \quad \text{and} \quad \hat{\theta}_{\text{FF}} = \frac{m_2(n - f_1) - f_1(\bar{x} - 1)^2}{n - \bar{x} f_1}\,. \tag{5.61}$$

Bardwell and Crow (1964) noted that parameter estimators obtained in this way are possibly better for the overdispersed case compared to the underdispersed one for the same value of θ. This is due to the size of frequency f_1, as evident in Figure 5.1.

size-biased HP model

The estimating equations for the size-biased HP model (5.23) are given by

$$\bar{x} = \theta - \lambda + 1 + \lambda \frac{{}_1F_1[1; \lambda; \theta]}{{}_1F_1[2; \lambda + 1; \theta]}\,, \quad \text{and} \quad \frac{f_1}{n} = \frac{1}{{}_1F_1[2; \lambda + 1; \theta]}\,. \tag{5.62}$$

Substituting ${}_1F_1[1; \lambda + 1; \theta] = \lambda\,({}_1F_1[1; \lambda; \theta] - 1)/\theta$ in relation (5.13) we have

$${}_1F_1[1; \lambda; \theta] = \theta\, \frac{{}_1F_1[2; \lambda + 1; \theta]}{\lambda(\theta - \lambda + 1)} + \frac{1 - \lambda}{\theta - \lambda + 1}\,. \tag{5.63}$$

The expression for ${}_1F_1[1; \lambda; \theta]$ obtained in this way is substituted in the first equation of (5.62) which by replacing ${}_1F_1[2; \lambda + 1; \theta]$ with n/f_1 results in

$$\bar{x} n(\theta - \lambda + 1) = n(\theta - \lambda + 1)^2 + n\theta + \lambda f_1(1 - \lambda)\,. \tag{5.64}$$

Ultimately, a few algebraic modifications enable us to rewrite it as following equation

$$\theta^2 - 2\theta\left(\lambda + \frac{\bar{x} - 3}{2}\right) + (\lambda - 1)\left(\lambda - 1 + \bar{x} - \frac{\lambda f_1}{n}\right) = 0\,. \tag{5.65}$$

(a) λ known: We can estimate θ from the quadratic equation above. Its positive root is the required estimator, which can be written as

$$\hat{\theta}_{\text{FF}} = \lambda + \frac{\bar{x} - 3}{2} + \sqrt{\left(\lambda + \frac{\bar{x} - 3}{2}\right)^2 - (\lambda - 1)\left(\lambda - 1 + \bar{x} - \frac{\lambda f_1}{n}\right)}\,. \tag{5.66}$$

(b) λ unknown: If there is no information about λ, derivation of the estimators $\hat{\lambda}_{\text{FF}}$ and $\hat{\theta}_{\text{FF}}$ needs an additional estimating equation. Analogously to the 1-displaced case, we use relation (5.27) and replace theoretical moments with the corresponding sample moments to obtain

$$m_2 = \theta(\bar{x}+1) + (\bar{x}-1)(1-\bar{x}-\lambda)\,. \tag{5.67}$$

Combining this equation with (5.65) and solving simultaneously the resulting system of equations we obtain the following quadratic equation in λ

$$\left(4 - \frac{(\bar{x}+1)^2 f_1}{n}\right)\lambda^2 - \left(4m_2 + (\bar{x}-3)^2 - \frac{(\bar{x}+1)^2 f_1}{n}\right)\lambda + c = 0 \tag{5.68}$$

where $c = \left(m_2 + (\bar{x}-1)^2\right)^2 - m_2(\bar{x}+1)(\bar{x}-3) - (\bar{x}^2-1)(\bar{x}^2-3\bar{x}+4)$. The estimator $\hat{\lambda}_{\text{FF}}$ is the positive root, whereas $\hat{\theta}_{\text{FF}}$ follows from (5.67) replacing λ with $\hat{\lambda}_{\text{FF}}$.

Moreover, Crow and Bardwell (1965) discussed further alternative approaches, as the one based on the mean and the first two frequencies. If the first two frequencies are much larger than the remaining ones, they recommended to use them exclusively to estimate parameters λ and θ. Another simple estimators could be obtained from the first three frequency classes where the second and third one are calculated from the recurrence relation for the successive HP probabilities. However, as emphasized by the authors, the usage of the frequencies is adequate only if they are “prominent features” of the distribution.

5.5 A Simulation Study

To investigate the behavior of the hyper-Poisson models, specified in (5.8) and (5.23) we performed a simulation study, reflecting the three possible dispersion situations: (i) overdispersion ($\delta > 1$), (ii) equidispersion ($\delta = 1$), (iii) underdispersion ($\delta < 1$). In order to be able to find out which of the two distributions might be preferable, we have to fix some of their common properties, otherwise comparison is not meaningful. Since both of them are two-parametric models we fix the first two moments, or equivalently the mean μ and the index of dispersion δ. The theoretical parameter pairs (λ, θ) corresponding to the selected μ and δ values (all precise to two decimal places) can be determined applying grid search. Table 5.4 exemplifies some of the possible parameter combinations. The sampling experiments are carried out to produce $M = 500$ Monte Carlo samples each of size $n = 500$ and $n = 1000$ for the parameter settings colored green in Table 5.4. These are identical to those given in the headings of Tables 5.5 and 5.6. The 1-displaced and size-biased HP random variables are generated by applying the inversion method, considered in Section 2.5, where the corresponding probabilities are computed using the recurrence relations (5.21) and (5.28), respectively. Tables 5.5 and 5.6 show the mean values of $M = 500$ parameter estimates $\hat{\lambda}$ and $\hat{\theta}$, denoted by $\bar{\lambda}$ and $\bar{\theta}$, for diverse dispersion situations and the three estimation techniques discussed in Section 5.4. Additionally, we calculate the estimated standard errors of the mean values as the standard deviation of the $M = 500$ parameter estimates. These are labelled by $\mathrm{se}_{\bar{\alpha}}$ and $\mathrm{se}_{\bar{\theta}}$ and measure the goodness of the resulting estimates. Although the obtained results are quite imprecise in both 1-displaced as well as size-biased HP model, the relative standard errors of the parameter estimates, calculated as $\mathrm{RSE}_{\bar{\lambda}} = \mathrm{se}_{\bar{\lambda}}/\bar{\lambda}$ and $\mathrm{RSE}_{\bar{\theta}} = \mathrm{se}_{\bar{\theta}}/\bar{\theta}$, show a clear trend among estimation techniques applied.

Table 5.4: Parameter settings for HP models when δ and μ are fixed at some value

			1-displaced		*size-biased*	
	δ	μ	λ	θ	λ	θ
$\delta > 1$	1.34	2.72	2.74	3.04	5.97	4.18
	1.27	**2.28**	**2.43**	**2.23**	**5.48**	**3.14**
	1.18	1.82	2.12	1.41	5.15	2.07
$\delta \approx 1$	**1.07**	**1.83**	**1.32**	**1.01**	**1.87**	**1.10**
	1.02	1.44	1.11	0.48	1.28	0.49
	0.99	1.25	0.87	0.22	0.65	0.21
$\delta < 1$	**0.90**	**1.73**	**0.65**	**0.54**	**0.12**	**0.47**
	0.85	2.64	0.57	1.28	0.11	1.17
	0.83	2.80	0.52	1.38	0.01	1.25

It is evident from Table 5.5 that in case of the 1-displaced HP model the smallest RSE are obtained for the ML method, followed by those arising from the FF approach, whereas the MM estimates have the largest RSE. Moreover, the increasing tendency of the relative standard errors is also apparent with respect to the various dispersion scenarios. We obtain the following relation: $\mathrm{RSE}_{(\delta>1)} < \mathrm{RSE}_{(\delta<1)} < \mathrm{RSE}_{(\delta\approx1)}$. Notice also that $\mathrm{RSE}_{\bar{\theta}} < \mathrm{RSE}_{\bar{\lambda}}$ holds, irrespective of the estimation technique and dispersion case.

Table 5.5: Estimation results for data simulated from the 1-displaced HP model

$\delta = 1.27$	$(\lambda, \theta) = (2.43, 2.23)$					
$\mu = 2.28$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	2.494 (0.890)	2.268 (0.549)	2.509 (0.676)	2.277 (0.435)	2.520 (0.683)	2.284 (0.439)
$n = 1000$	2.457 (0.621)	2.245 (0.383)	2.453 (0.449)	2.243 (0.288)	2.459 (0.453)	2.246 (0.292)
$\delta = 1.07$	$(\lambda, \theta) = (1.32, 1.01)$					
$\mu = 1.83$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	1.364 (0.560)	1.029 (0.287)	1.378 (0.402)	1.036 (0.220)	1.382 (0.408)	1.039 (0.223)
$n = 1000$	1.337 (0.387)	1.018 (0.199)	1.337 (0.275)	1.018 (0.151)	1.341 (0.281)	1.020 (0.154)
$\delta = 0.9$	$(\lambda, \theta) = (0.65, 0.54)$					
$\mu = 1.73$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.657 (0.278)	0.544 (0.142)	0.663 (0.175)	0.547 (0.103)	0.666 (0.188)	0.549 (0.109)
$n = 1000$	0.653 (0.195)	0.541 (0.098)	0.653 (0.117)	0.541 (0.069)	0.656 (0.127)	0.543 (0.073)

Figure 5.3 illustrates the degree of dependence between parameter estimates of λ and θ for $M = 500$ generated 1-displaced HP samples of size $n = 1000$ for each of the relevant estimation methods and the three dispersions situations. The estimated

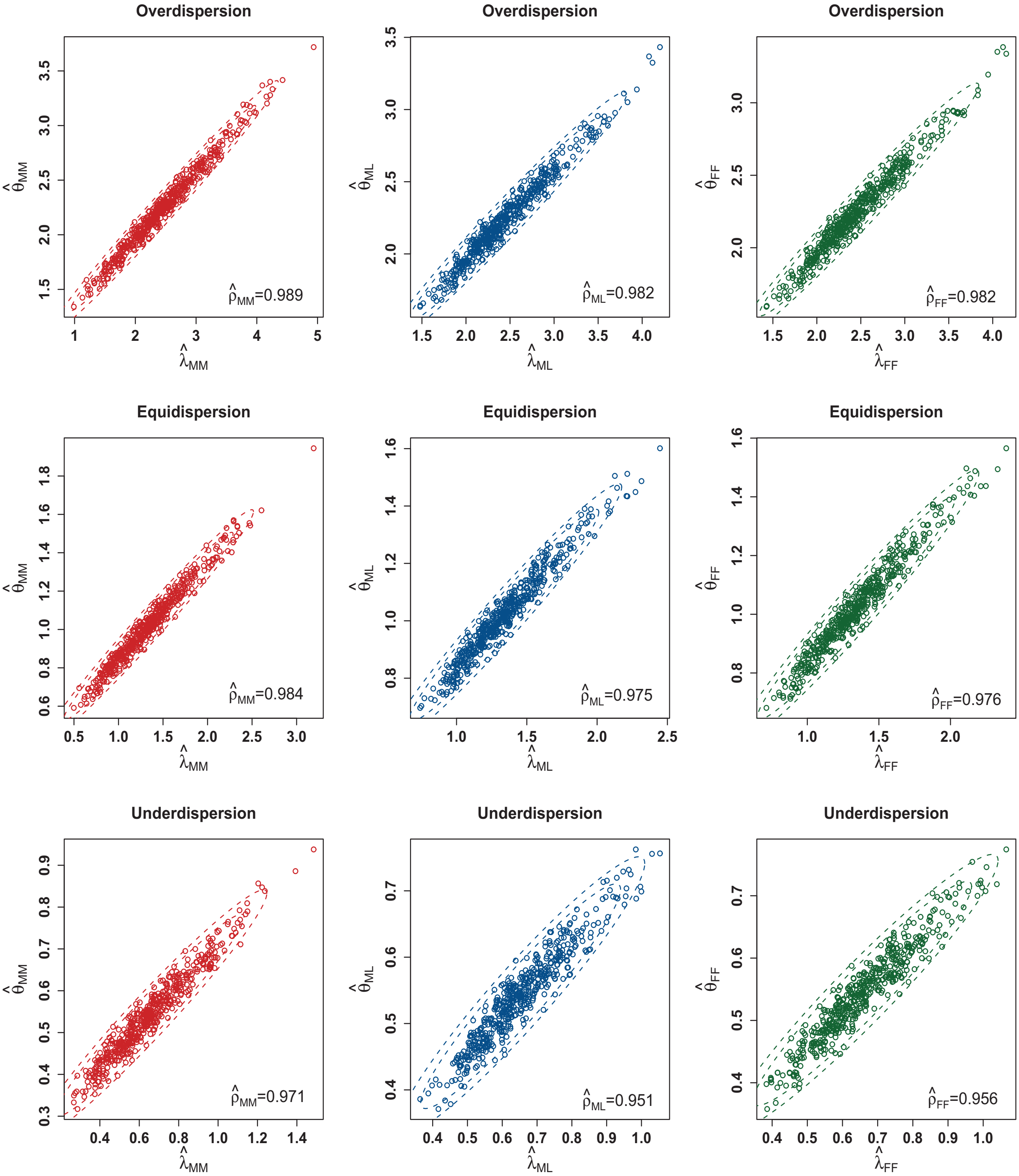


Figure 5.3: Correlation between parameters of the 1-displaced HP model for $M = 500$ created Monte Carlo samples of size $n = 1000$ in over-, equi and underdispersed case.

correlation coefficient of $M = 500$ replications is also reported. Additionally, the 95% and 99% confidence ellipses of the corresponding parameter pairs are drawn.

Table 5.6 shows the results of the simulation experiment for the size-biased HP model. As to the method based on the first frequency class (FF) we found out that the parameter estimates are considerably biased compared to the other two methods, except for the underdispersed case in which for all three techniques we obtained comparable results. Apparently, here we also have $\mathrm{RSE}_{\bar{\theta}} < \mathrm{RSE}_{\bar{\lambda}}$. The greatest accuracy have the estimates of θ when $\delta < 1$. However, in this case the most imprecise results for λ are obtained. It should also be noticed that in some cases we could not get all $M = 500$ estimation results. The number of the effective samples providing reliable estimates are given in Table 5.7 for each of the three dispersions scenarios and relevant estimation methods.

Table 5.6: Estimation results for data simulated from the size-biased HP model

$\delta = 1.27$	$(\lambda, \theta) = (5.48, 3.14)$					
$\mu = 2.28$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	6.076 (3.758)	3.371 (1.485)	6.234 (3.560)	3.434 (1.413)	7.233 (6.461)	3.826 (2.511)
$n = 1000$	5.813 (2.227)	3.271 (0.898)	5.882 (2.064)	3.298 (0.834)	6.119 (2.687)	3.391 (1.076)
$\delta = 1.07$	$(\lambda, \theta) = (1.87, 1.10)$					
$\mu = 1.83$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	2.119 (1.629)	1.171 (0.488)	2.177 (1.416)	1.189 (0.428)	2.554 (1.615)	1.301 (0.481)
$n = 1000$	2.022 (1.026)	1.145 (0.314)	2.051 (0.904)	1.153 (0.277)	2.315 (0.888)	1.232 (0.272)
$\delta = 0.90$	$(\lambda, \theta) = (0.12, 0.47)$					
$\mu = 1.73$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.455 (0.414)	0.569 (0.120)	0.366 (0.327)	0.544 (0.101)	0.345 (0.286)	0.538 (0.089)
$n = 1000$	0.323 (0.288)	0.531 (0.087)	0.284 (0.228)	0.520 (0.072)	0.278 (0.223)	0.519 (0.071)

Table 5.7: Number of samples (out of M=500) providing valid estimates for the three estimating methods and various dispersion cases

	method of moments			*maximum likelihood*			*first frequency class*		
Sample size	$\delta > 1$	$\delta \approx 1$	$\delta < 1$	$\delta > 1$	$\delta \approx 1$	$\delta < 1$	$\delta > 1$	$\delta \approx 1$	$\delta < 1$
$n = 500$	500	496	273	500	499	322	487	428	316
$n = 1000$	500	500	328	500	500	346	499	405	351

Figure 5.4 shows very high correlation between parameter estimates $\hat{\lambda}$ and $\hat{\theta}$ for the $M = 500$ size-biased HP samples of size $n = 1000$ irrespective of the estimation method and dispersion situations.

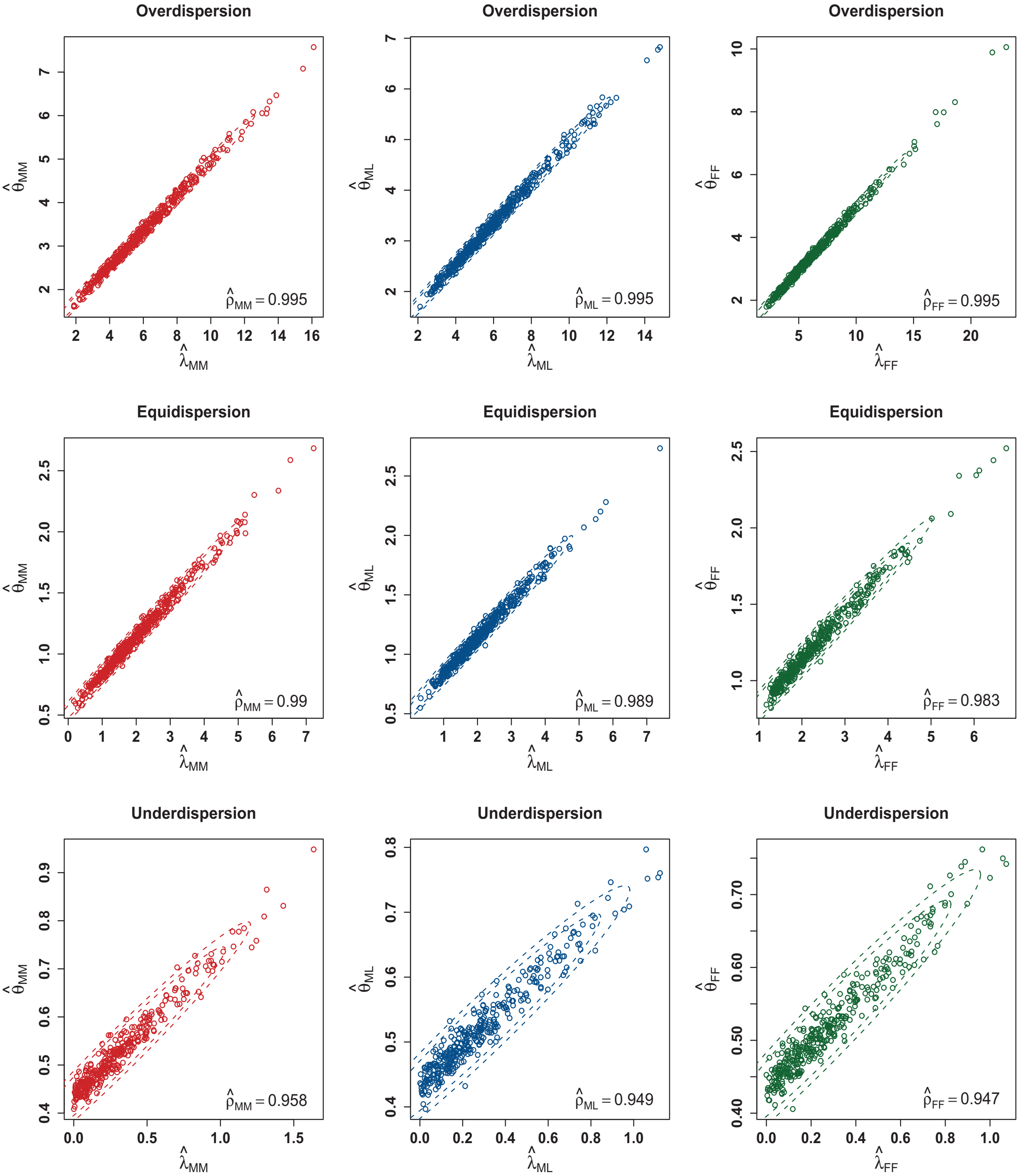


Figure 5.4: Correlation between parameters of the size-biased HP model for $M = 500$ created Monte Carlo samples of size $n = 1000$ in over-, equi- and underdispersed case.

Chapter 6

Generalized Poisson Distribution

6.1 Introduction

The generalized Poisson (GP) distribution was first introduced by Consul and Jain (1973a) as a limiting form of the generalized negative binomial distribution. Hence, it is also known in the literature as the *Consul's generalized Poisson distribution.* Moreover, this distribution can be derived as a Poisson-stopped sum of N Borel random variables with the pmf defined by (cf. Consul and Famoye, 2006, p. 158)

$$P(Y=y)=\frac{(\lambda y)^{y-1}e^{-\lambda y}}{y!}\,, \quad y=1,2,\dots\,, \tag{6.1}$$

and zero otherwise, where $0<\lambda<1$. The Borel distribution belongs to the *basic Lagrangian distributions*, and has the pgf defined by $G_Y(t)=z$, where $z=te^{\lambda(z-1)}$. The Lagrangian distributions[1] have been extensively studied by Consul and Famoye (2006) and include numerous useful discrete distributions applicable in various fields. Since the pgf of $N\sim$ Poisson(θ) is $G_N(t)=e^{\theta(t-1)}$, then applying formula (2.24), Section 2.2.2, the pgf of the random variable $X=Y_1+Y_2+\ldots+Y_N$ is given by

$$G_X(t)=G_N(G_Y(t))=G_N(z)=e^{\theta(z-1)}\,, \quad \text{where} \quad z=te^{\lambda(z-1)}\,. \tag{6.2}$$

Using Gurland's notation we can represent symbolically this process of generalizing[2] the Poisson distribution by the Borel distribution as $X\sim \text{Poisson}(\theta)\bigvee \text{Borel}(\lambda)$. The generalized Poisson distribution considered here is an important member of the class of *general Lagrangian distributions*, formed from the functions $f(z)=e^{\theta(z-1)}$ and $g(z)=e^{\lambda(z-1)}$ that satisfy the necessary conditions $g(0)\neq 0$, $g(1)=1$, $f(1)=1$ and

$$\left[\frac{\partial^{x-1}}{\partial z^{x-1}}\left((g(z))^x\frac{\partial f(z)}{\partial z}\right)\right]_{z=0}\geq 0\,, \ \text{for} \ x\in\mathbb{N}\,,$$

[1] Discrete Lagrangian probability distributions can, according to their probabilistic structure, be divided into three subclasses: (i) basic, (ii) delta and (iii) general Lagrangian distributions.

[2] For the preliminaries regarding the generalized distributions see Section 2.2.2.

for generating Lagrangian distributions (cf. Consul and Famoye, 2006, p. 165). Thus, the pgf of X can be obtained by the Lagrangian expansion of $f(z)$ in powers of t, under the transformation $z = tg(z)$, as follow (cf. Consul and Famoye, 2006, p. 25)

$$G_X(t) = f(z) = \sum_{x=0}^{\infty} \frac{t^x}{x!} \left[\frac{\partial^{x-1}}{\partial z^{x-1}} \left((g(z))^x \, \frac{\partial f(z)}{\partial z} \right) \right]_{z=0} , \quad x \in \mathbb{N} . \tag{6.3}$$

The pmf of the GP distribution is then given by the following formula

$$P(X=0) = f(0) , \quad P(X=x) = \frac{1}{x!} \left[\frac{\partial^{x-1}}{\partial z^{x-1}} \left((g(z))^x \, \frac{\partial f(z)}{\partial z} \right) \right]_{z=0} , \quad x \in \mathbb{N} , \tag{6.4}$$

and for $\theta > 0$ being independent of $0 < \lambda < 1$ it results in (see also Consul, 1989)

$$\pi_{x|\lambda,\theta} = P(X=x) = \frac{\theta(\theta + x\lambda)^{x-1} e^{-(\theta + x\lambda)}}{x!} , \quad x = 0, 1, \ldots \tag{6.5}$$

Notice that in case when $\lambda < 0$ the term $(\theta + x\lambda)$ becomes negative for $\lambda < -\theta/x$, $x \neq 0$, thus the values in (6.5) get also positive or negative depending on whether the power $(x-1)$ is an even or odd number, respectively. To avoid this difficulty Consul and Jain (1973a) suggested to bring into use truncation of the model to $m+1$ mass points, i.e. to adopt $\pi_{x|\lambda,\theta} = 0$ for all $x > m$, where m denotes the largest positive integer x for which $\theta + x\lambda > 0$. This kind of a non-standard truncation causes for some particular parameter values that the total sum of probabilities in (6.5) differs from unity. For illustration, we assume that $\theta = 1.35$, $\lambda = -0.65$ and $m = 2$. We get $\pi_{0|\lambda,\theta} = 0.2592$, $\pi_{1|\lambda,\theta} = 0.6704$, $\pi_{2|\lambda,\theta} = 0.0321$ and $\pi_{x|\lambda,\theta} = 0$ for all $x > 2$, but $\sum \pi_{x|\lambda,\theta} = 0.9617$. The solution of this problem, as pointed out by Consul (1989), is to normalize the pmf in (6.5) by dividing each probability with $\sum_{x=0}^{m} \pi_{x|\lambda,\theta}$ whenever $\lambda < 0$. However, Consul and Shoukri (1985) made detailed error analysis showing that in such a case one should rather consider the estimated value of θ and the number of non-zero probability classes $m+1$, since the error of truncation may be serious only when simultaneously holds $m \in \{2, 3\}$ and the estimated value of θ is bigger than 1.5 (see also Consul, 1989, Section 9.1.1). In all other cases where $\lambda < 0$ the model (6.5) can be applied without any additional correction for truncation. In the next section we show that negative values of λ indicate in fact $\bar{x} > s^2$ data cases.

The determination of the mean and variance of the GP distribution by direct summation is pretty complex since the parameters λ and θ appear in (6.5) in linear functional form. Using a tricky method instead, Consul and Jain (1973a) got simple expressions for the first two moments given in Table 6.1. Next section proves that based on the first two moments of the Borel distribution, however, their derivation is rather simple.

The GP model has found its successful application in various practical situations in which the inequality of the mean and variance is present, and the observed counts exhibit either underdispersion or overdispersion. Moreover, it proved to be useful for fitting equidispersed data, and thus contains a Poisson model as its special case.

Neubauer and Djuraš (2008) considered this model in regression context to estimate the total number of the crimes committed. However, since certain criminal activities, such as those associated with shame (e.g. sexual offence, domestic violence), theft of low value goods or victims being themselves criminals, are likely not to be reported to the police, we have to deal with a problem of underreporting. The determination of the number of unreported occurrences, the so-called *"dark figures"*, out of the expected total number of cases and available counts is of enormous importance (cf. Neubauer, Djuraš, and Friedl, 2010). Moreover, in public health there are register systems for infectious and chronic diseases, and the recording errors often occur as a result of misdiagnosis or patients avoiding medical screening. In this case it is also beneficial to figure out the true number of the sick people in the population. Numerous further examples including home injuries, behavior patterns of bacteria in the human body or pest eggs on the plant leaves, traffic research studies of the number of cars travelling the street, etc., all being not Poisson distributed could be successfully described by the GP distribution, as referred by Consul (1989). Based on simulation studies Neubauer, Djuraš, and Friedl (2009) concluded that for $0 < \lambda < 1$ the GP model is near to the negative binomial (NB) model, and whenever λ is negative it nearly corresponds to the binomial model. The parameter λ provides therefore an information about the type of the distribution, and the parameter θ indicates the intensity of the natural Poisson process. Figure 6.1 illustrates differences in the pmf between the GP and NB models for $s^2 > \bar{x}$, whereas probabilistic comparison for the GP and binomial (B) distributions when $s^2 < \bar{x}$ is shown in Figure 6.2. In both cases the first two moments are fixed. As a consequence, we have the following parametrization of the model parameters according to the mean and variance

$$\text{for NB}(r,p): \ \mu = \frac{(1-p)r}{p}\,, \ \ \sigma^2 = \frac{(1-p)r}{p^2}\,, \ \text{thus } r = \frac{\mu^2}{\sigma^2-\mu} \ \text{ and } \ p = \frac{\mu}{\sigma^2}\,,$$

$$\text{for GP}(\lambda,\theta): \ \mu = \frac{\theta}{1-\lambda}\,, \ \ \sigma^2 = \frac{\theta}{(1-\lambda)^3}\,, \ \text{thus } \lambda = 1-\sqrt{\frac{\mu}{\sigma^2}} \ \text{ and } \ \theta = \mu\sqrt{\frac{\mu}{\sigma^2}}\,,$$

$$\text{for B}(n,p^*): \ \mu = np^*\,, \ \ \sigma^2 = np^*(1-p^*)\,, \ \text{thus } n = \frac{\mu^2}{\mu-\sigma^2} \ \text{ and } \ p^* = \frac{\mu-\sigma^2}{\mu}\,.$$

In Figure 6.1 the mean is taken to be $\mu = 0.5, 1.5, 5, 15$ and the index of dispersion $\delta = 1.5, 5, 10$ in order to demonstrate low, large and very large overdispersion cases. Each row contains three plots having the same value of μ, and in each column we have those four with the same value of δ. The rows thus indicate if there is any difference in the probability distribution that results from increase in dispersion, whereas the columns describe modifications according to the change in mean. The pmf of the GP distribution is computed by using the ratio of two successive probabilities, suggested by Consul (1989, p. 19). Obviously, for low overdispersion (when δ is close to 1) there is negligible difference in the pmf's. The same happens for $\mu = 0.5$, irrespective of the value of δ. When μ becomes much bigger, the pmf's differ as δ increases. However, the difference becomes significant only for very large values of δ. Interestingly, the

NB distribution has larger zero probability compared to that of the GP model, which is even more apparent for bigger values of δ. For this reason someone might prefer to use the NB model instead of the GP model when the observed counts exhibit to many zeros relative to the Poisson case. Nevertheless, when sample data manifest jointly the excess of zero observations and heavy right tails, it is beneficial to use the zero-inflated GP distribution to get a better fit (cf. Joe and Zhu, 2005; Famoye and Singh, 2006).

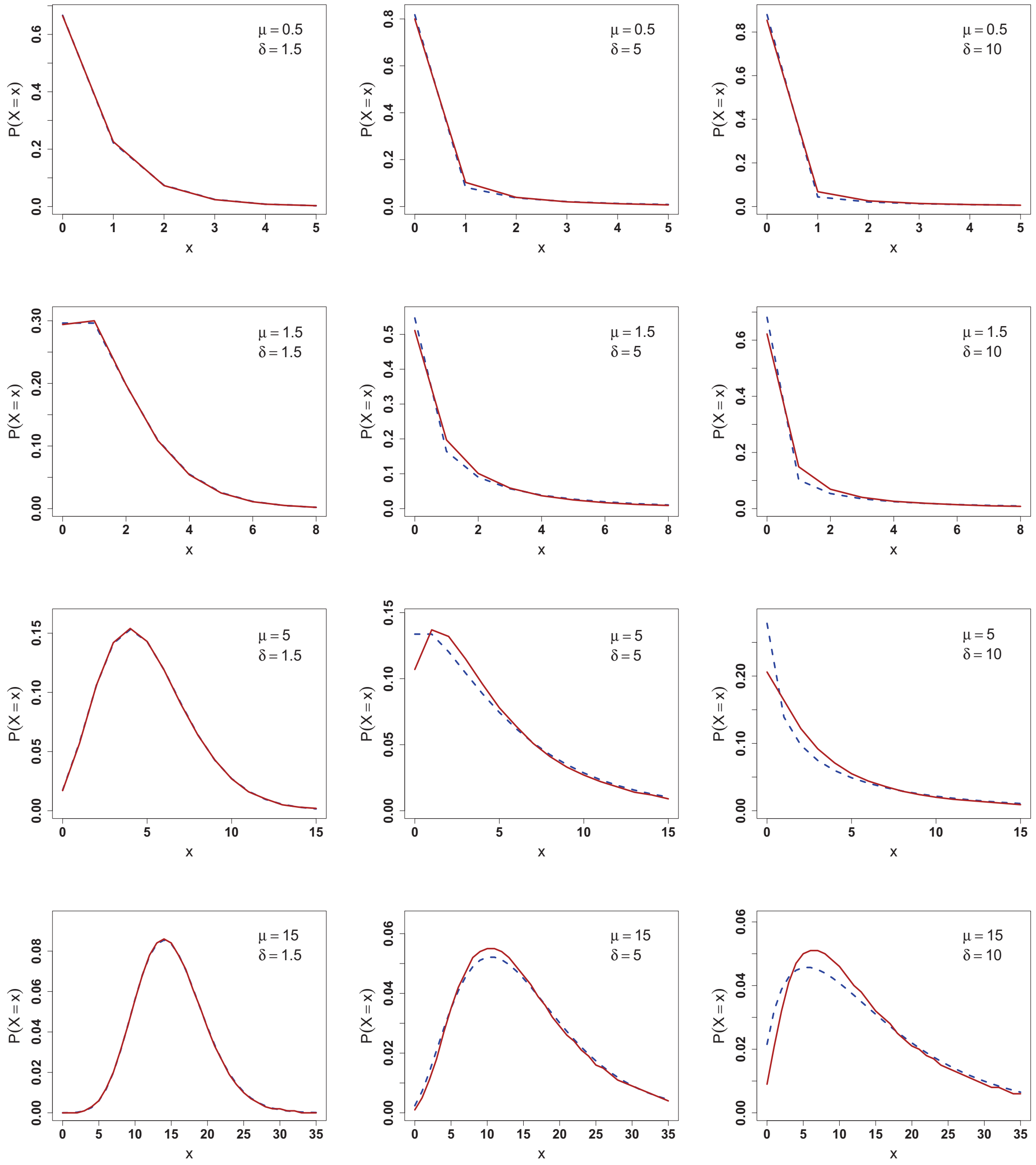


Figure 6.1: Differences in pmf's between GP(λ, θ) (red solid line) and NB(r, p) (blue dashed line) for mean $\mu = 0.5, 1.5, 5, 15$ and $\delta = 1.5, 5, 10$ overdispersion cases.

Figure 6.2 contains 12 graphs for $\mu = 1.5, 7.5, 16.5$, and $\delta = 0.25, 0.5, 0.75, 0.95$, to visualize various underdispersion cases. Each column shows plots with the same value of μ, and rows display those with the same indices of dispersion. The difference between the binomial and the corresponding GP distribution is so small that the two lines overlap in most cases. The slight disagreement is only seen for $\delta = 0.25$, but vanishes as μ increases. To sum up, the GP model approximates both the negative binomial and binomial model, being thus suitable for both types of dispersion.

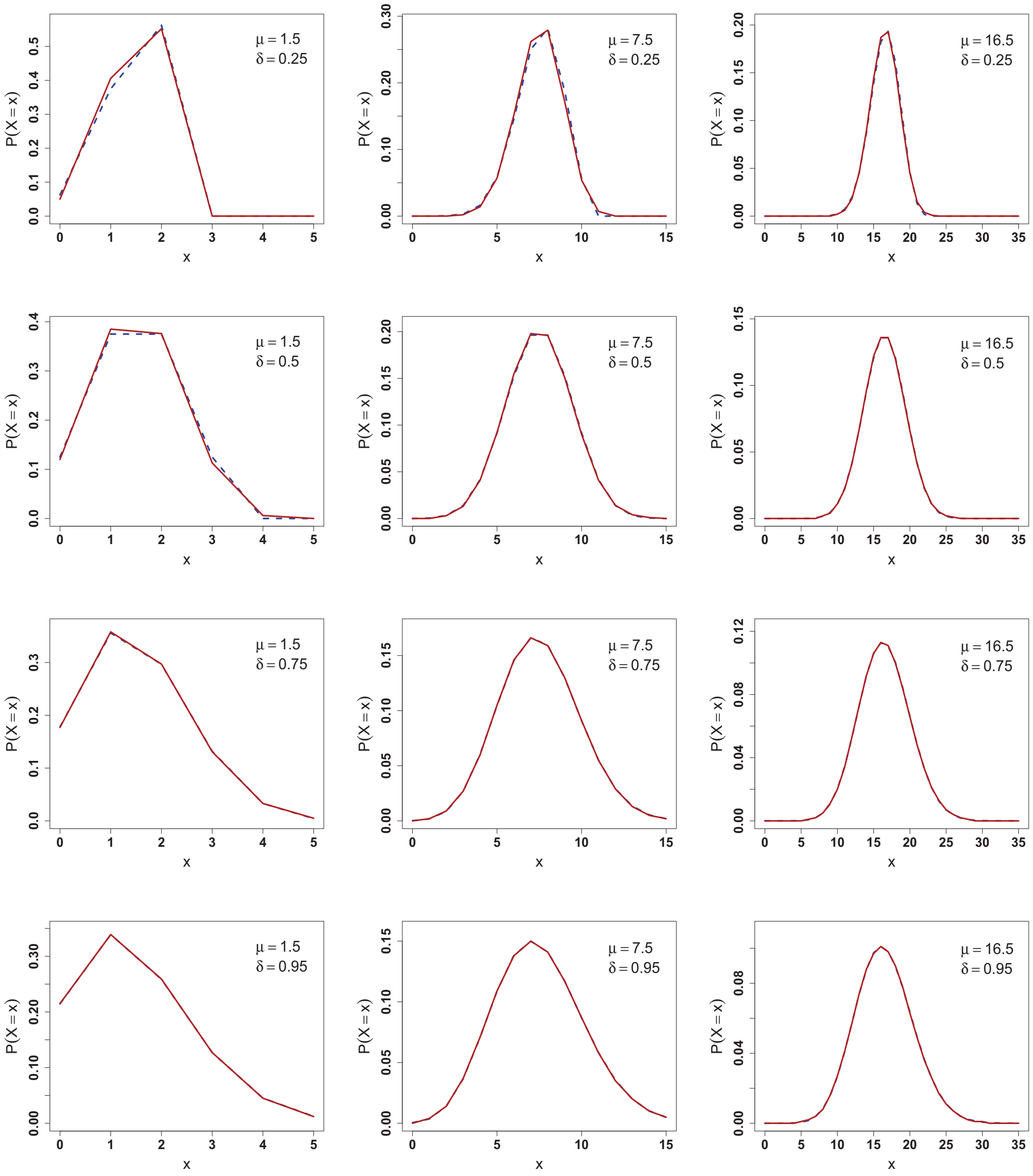


Figure 6.2: Differences in pmf's between GP(λ, θ) (red solid line) and B(n, p^*) (blue dashed line) for mean $\mu = 0.5, 1.5, 5, 15$ and $\delta = 0.25$, 0.5, 0.75, 0.95 underdispersion cases.

It should be also noted that the GP distribution possesses an important property that facilitates its wide application in practice. Namely, it is closed under addition, as mathematically formulated by the following theorem (cf. Consul and Jain, 1973b).

Theorem 6.1 *The sum $X_1 + X_2$ of two independent GP random variables X_1 and X_2, with parameters (λ, θ_1) and (λ, θ_2), respectively, is itself a GP random variable with parameters $(\lambda, \theta_1 + \theta_2)$.*

This theorem holds even more generally. The sum of n independent random variables $X_i \sim GP(\lambda, \theta_i)$, is again a GP random variable, where $\sum_{i=1}^{n} X_i \sim GP(\lambda, \sum \theta_i)$. Therefore, we can obtain equivalent results for individual and grouped data provided that at least the parameter λ stays fixed in all groups.

Table 6.1 compares the main features of the original GP distribution to those of its 1-displaced and size-biased versions, both discussed in the forthcoming sections.

Table 6.1: Generalized Poisson distribution: original, 1-displaced and size-biased forms

	Random Variable		
Distribution	X	X^d	X^*
Notation	GP(λ, θ)	1+GP(λ, θ)	sbGP(λ, θ)
Range	$\mathbb{N}_0$	$\mathbb{N}$	$\mathbb{N}$
pmf	$\dfrac{\theta(\theta + x\lambda)^{x-1} e^{-(\theta+x\lambda)}}{x!}$	$\pi_{x-1\mid\lambda,\theta}$	$\dfrac{(1-\lambda)x}{\theta}\pi_{x\mid\lambda,\theta}$
pgf	as in (6.2)	as in (6.7)	as in (6.15)
E$(\cdot)$	$\theta/(1-\lambda)$	E$(X) + 1$	E$(X) + 1/(1-\lambda)^2$
var$(\cdot)$	$\theta/(1-\lambda)^3$	var(X)	var$(X) + 2\lambda/(1-\lambda)^4$

6.2 1-Displaced Generalized Poisson Distribution

Suppose X is a discrete random variable with pmf $\pi_{x\mid\lambda,\theta}$ given in (6.5). The pmf corresponding to the distribution of the linear transformation $X^d = X + 1$ defined only over positive integers is then given by the formula

$$\pi^d_{x\mid\lambda,\theta} = P(X^d = x) = \begin{cases} \dfrac{\theta(\theta + x\lambda - \lambda)^{x-2} e^{-(\theta+x\lambda-\lambda)}}{(x-1)!}, & x = 1, \ldots, m+1\,, \\ 0, & x > (m+1), \text{ for } \lambda < 0 \end{cases} \tag{6.6}$$

and zero otherwise, where $\theta > 0$, $\max(-1, -\theta/m) \leq \lambda < 1$ and m is the largest positive integer such that $\theta + m\lambda > 0$ when λ is negative. Additionally, the condition

$m \geq 4$ is proposed in order to ensure that there are at least five non-zero probability classes in the truncated model when $\lambda < 0$ (cf. Consul, 1989, p. 4). Notice that in the case when $0 \leq \lambda < 1$, the support of the model (6.6) needs not to be truncated, hence we have $m = \infty$. When λ is negative we obtain $\pi^d_{x|\lambda,\theta} > 0$ only for $x \leq (m+1)$, but m depends now on the unknown parameters θ and λ. Apparently, for $\lambda = 0$ the distribution above reduces to the 1-displaced Poisson distribution with parameter θ.

The pgf of X^d can be obtained from that of X, given in (6.2), and has again the implicit form defined by

$$G_{X^d}(t) = tG_X(t) = te^{\theta(z-1)}\,, \quad \text{where} \quad z = te^{\lambda(z-1)}\,. \tag{6.7}$$

To derive the mean and variance of X^d we return to the fact that the GP distribution is a Poisson-stopped sum of N Borel random variables Y_i, defined by (6.1), with mean $\mathrm{E}(Y) = 1/(1-\lambda)$ and variance $\mathrm{var}(Y) = \lambda/(1-\lambda)^3$ (cf. Consul and Famoye, 2006). Since $N \sim \text{Poisson}(\theta)$, by using relation (2.32), Section 2.2.2 we get for the first two moments of $X = Y_1 + \ldots + Y_N \sim \mathrm{GP}(\lambda, \theta)$ the following expressions

$$\mathrm{E}(X) = \theta\mathrm{E}(Y) = \frac{\theta}{1-\lambda} \quad \text{and} \quad \mathrm{var}(X) = \theta\mathrm{E}(Y^2) = \frac{\theta}{(1-\lambda)^3}\,. \tag{6.8}$$

Consequently, we obtain the mean and variance of the distribution in (6.6) as

$$\mathrm{E}(X^d) = \mathrm{E}(X) + 1 = \frac{\theta}{1-\lambda} + 1 \quad \text{and} \quad \mathrm{var}(X^d) = \mathrm{var}(X) = \frac{\theta}{(1-\lambda)^3}\,, \tag{6.9}$$

and therefore the index of dispersion takes the simple form

$$\delta = \frac{\mathrm{var}(X^d)}{\mathrm{E}(X^d) - 1} = \frac{1}{(1-\lambda)^2}\,. \tag{6.10}$$

Accordingly, the parameter λ has an important role as a diagnostic measure, since it determines the dispersion in model (6.6). Obviously, $\lambda = 0$ indicates the presence of equality of mean and variance (i.e. equidispersion case), and the GP in (6.6) reduces to the common Poisson model. For $0 < \lambda < 1$ we have $\delta > 1$, the feature of overdispersion, hence the model here allows for modelling counts having $s^2 > \bar{x} - 1$. When $\lambda < 0$ it follows that $\delta < 1$, and the distribution (6.6) becomes underdispersed. Therefore, it also enables to describe $s^2 < \bar{x} - 1$ data cases.

To derive the factorial moments of the distribution above we combine the relation (2.38), Section 2.3.1 with the formula (B.8), Appendix B, and obtain

$$\mu^d_{(k)} = \sum_{i=0}^{k} s(k,i)\mu'_i + k\sum_{i=0}^{k-1} s(k-1,i)\mu'_i\,, \tag{6.11}$$

where $s(k,i)$ are the Stirling numbers of the first kind specified in Appendix B, and μ'_i is the i-th raw moment of the original GP distribution (6.5), defined in Consul (1989, p. 50) by the following recurrence formula

$$(1-\lambda)\mu'_{k+1} = \theta\mu'_k + \theta\frac{\partial\mu'_k}{\partial\theta} + \lambda\frac{\partial\mu'_k}{\partial\lambda}\,, \quad k = 0, 1, \ldots \tag{6.12}$$

With the factorial moments determined by (6.11), all central and raw moments of distribution (6.6) exist for $\theta > 0$ and $\lambda < 1$, and can be found using appropriate formulas from Appendix B.

Figures 6.3 and 6.4 display differences in the probability mass functions between the 1-displaced GP distribution and its size-biased competitor with the pmf (6.14) defined in the next section, for positive and negative values of λ, respectively. The individual probabilities of the GP distribution (6.6) are computed using the following ratio of successive probabilities

$$\frac{\pi^d_{x|\lambda,\theta}}{\pi^d_{x-1|\lambda,\theta}} = \frac{(\theta + x\lambda - \lambda)^{x-2} e^{-\lambda}}{(x-1)(\theta + x\lambda - 2\lambda)^{x-3}}\,, \quad \text{for } x \geq 2\,, \tag{6.13}$$

where $\pi^d_{1|\lambda,\theta} = e^{-\theta}$, as given in (6.6) for $x = 1$. For the calculation of the corresponding size-biased probabilities we apply the recurrence relation (6.19). To illustrate various overdispersion cases we choose λ to be 0.1, 0.2, 0.4, and 0.6 (cf. Figure 6.3), whereas to allow for different degrees of underdispersion we take negative values of λ, in fact -0.1, -0.2, -0.4, and -0.6 (cf. Figure 6.4). Subplots in each row possess the same value of λ and highlight alternations in the pmf caused by the change in the value of θ. Columns, however, have the same value of θ, and thus clearly show the effect of varying the value of λ on the shape of the distribution. Since only the value of θ participates in the computation of the first probability, changes in the mass at unity are visible exclusively between columns.

From both figures we can see that the 1-displaced GP distribution is L-shaped for small values of θ, though becomes bell-shaped as θ increases. Moreover, its profile is more L-shaped for $0 < \lambda < 1$ and less L-shaped for the negative values of λ. For the values of λ close to zero the plot of the 1-displaced GP model differs only slightly from the 1-displaced Poisson case, however the difference becomes larger as λ increases or decreases. It can be also seen that an increase in the value of λ by the fixed value of θ, increases the mean and variance of the distribution as well (cf. formula (6.9)). The bell-shaped form of the distribution then becomes flatter and the cupola of the bell lower, as clearly evident from the rightmost columns. Note that for the negative values of λ the right-hand tail of the distribution becomes in this case longer. Furthermore, when $\lambda < 0$ we have less mass points due to the fact that for all values of x for which $\theta + x\lambda < 0$ the corresponding probabilities $\pi^d_{x|\lambda,\theta}$ are zero. Choosing e.g. $\lambda = -0.6$ and $\theta = 1.5$ it follows that $m = 2$, thus we have three non-zero probability classes, i.e. $\pi^d_{x|\lambda,\theta} = 0$ for $x > 3$, as the plot in the middle of the last row in Figure 6.4 correctly displays. The behavior of the size-biased GP distribution is similar to that of the 1-displaced GP distribution. However, we see that $\pi^d_{1|\lambda,\theta} > \pi^*_{1|\lambda,\theta}$ holds for the positive values of λ, while for the negative values of λ we have $\pi^d_{1|\lambda,\theta} < \pi^*_{1|\lambda,\theta}$. These differences become even bigger, when λ increases in absolute value. We will see in the next section that this is due to the multiplication factor that is dependent on parameter λ in the size-biased case.

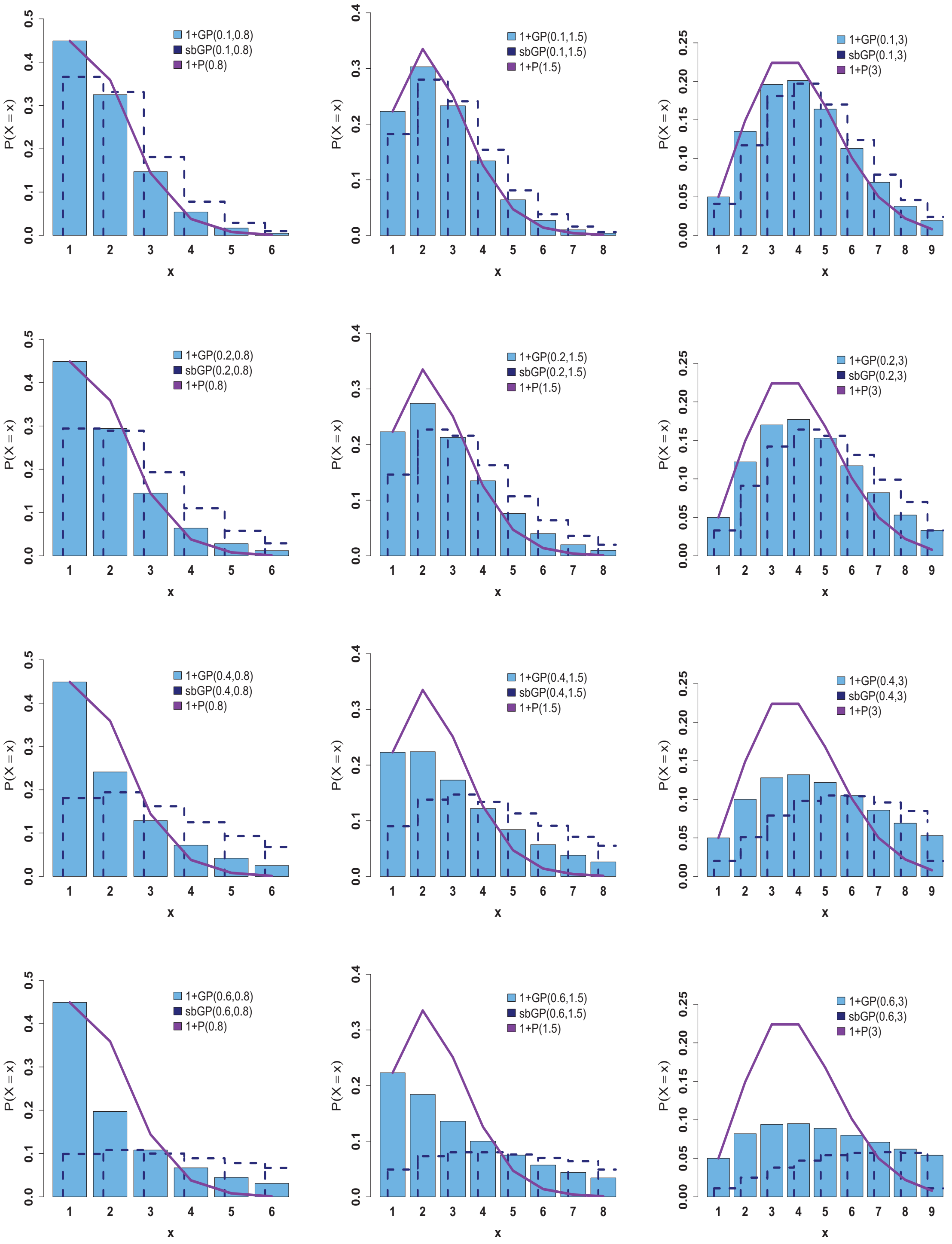


Figure 6.3: Differences in 1-displaced GP (1+GP(λ, θ)), size-biased GP (sbGP(λ, θ)) and Poisson (1+P(θ)) distributions for $\theta = 0.8$, 1.5, 3, $\lambda = 0.1$, 0.2, 0.4, 0.6 and $\delta > 1$.

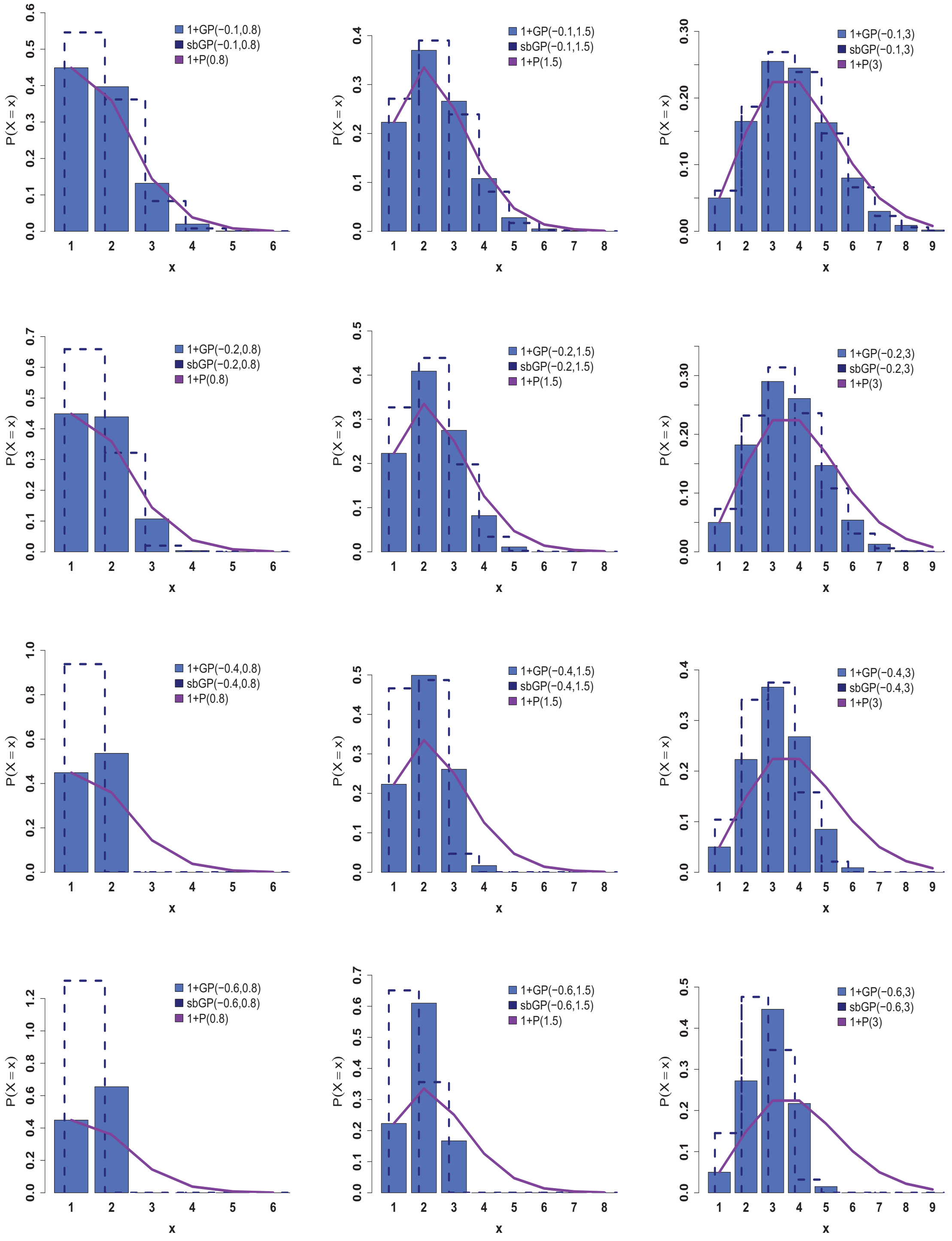


Figure 6.4: Differences in 1-displaced GP (1+GP(λ, θ)), size-biased GP (sbGP(λ, θ)) and Poisson (1+P(θ)) distributions for $\theta = 0.8, 1.5, 3$, $\lambda =$ -0.1, -0.2, -0.4, -0.6 and $\delta < 1$.

6.3 Size-Biased Generalized Poisson Distribution

To obtain the size-biased GP distribution we apply the transformation (2.41), Section 2.3.2 to the pmf given in (6.5), with $\mathrm{E}(X)$ defined by the first expression of (6.8). Then, the pmf of the size-biased GP random variable X^* results in

$$\pi^*_{x|\lambda,\theta} = P(X^* = x) = \begin{cases} \dfrac{(1-\lambda)(\theta + x\lambda)^{x-1} e^{-(\theta + x\lambda)}}{(x-1)!}\,, & x = 1, \ldots, m\,, \\ 0 \,, & x > m\,, \text{ for } \lambda < 0\,, \end{cases} \tag{6.14}$$

where $\theta > 0$ and all necessary conditions regarding the domain of the parameter λ mentioned before hold also here. For $\lambda < 0$ the support of (6.14) depends on the unknown parameters θ and λ, as the process of determining m requires both of them. Clearly, for $\lambda = 0$ the above distribution simplifies to the 1-displaced Poisson.

The pgf of X^* can be found by the relation (2.42), Section 2.3.2, which, however, requires derivative of the implicit pgf in (6.2). Denoting $z = z(t) = te^{\lambda(z(t)-1)}$ and taking its derivative with respect to t we have the following relation

$$z'(t) = e^{\lambda(z(t)-1)} + te^{\lambda(z(t)-1)}\lambda z'(t)\,,$$

which after rearranging leads to the following equality

$$z'(t) = \frac{e^{\lambda(z(t)-1)}}{1 - t\lambda e^{\lambda(z(t)-1)}} = \frac{1}{e^{-\lambda(z(t)-1)} - t\lambda}\,.$$

Finally, the pgf of X^* becomes

$$G_{X^*}(t) = \frac{t(1-\lambda)e^{\theta(z-1)}}{e^{-\lambda(z-1)} - t\lambda}\,, \quad \text{where} \quad z = te^{\lambda(z-1)}\,. \tag{6.15}$$

All moments of the size-biased GP model exist for $\theta > 0$ and $\lambda < 1$. Applying relation (2.46), Section 2.3.2, the k-th size-biased raw moment is obtained as a ratio of the $(k+1)$-th raw moment, denoted by μ'_{k+1}, and the mean of the unmodified model (6.5). These are given by (6.12) and the first equation of (6.8), respectively. Thus, the raw moment of distribution (6.14) follows from the expression below

$$\mu^{*\prime}_k = \mu'_{k+1}(1-\lambda)\theta^{-1}\,, \;\; k = 1, 2, \ldots \tag{6.16}$$

With the raw moments calculated in this way, all factorial and central moments of X^* are obtained by using the relations (B.8) and (B.13), Appendix B, respectively. As a consequence of (6.16), we can write the mean and variance of (6.14) as follows

$$\mathrm{E}(X^*) = \mu'_2(1-\lambda)\theta^{-1} \;\text{ and }\; \mathrm{var}(X^*) = \mu'_3(1-\lambda)\theta^{-1} - (\mathrm{E}(X^*))^2\,,$$

which by substituting the values of μ'_2 and μ'_3 obtained from (6.12) finally results in

$$\begin{aligned} \mathrm{E}(X^*) &= \theta(1-\lambda)^{-1} + (1-\lambda)^{-2} = (\theta(1-\lambda)+1)\,(1-\lambda)^{-2}\,, \\ \mathrm{var}(X^*) &= \theta(1-\lambda)^{-3} + 2\lambda(1-\lambda)^{-4} = (\theta(1-\lambda)+2\lambda)\,(1-\lambda)^{-4}. \end{aligned} \tag{6.17}$$

In contrast to (6.10), the index of dispersion has here a more complex form given by

$$\delta = \frac{\mathrm{var}(X^*)}{\mathrm{E}(X^*) - 1} = \frac{\theta(1-\lambda) + 2\lambda}{(1-\lambda)^2\,(1 + (1-\lambda)(\theta + \lambda - 1))}\,. \tag{6.18}$$

From this expression we can not easily conclude whether the distribution (6.14) has capability for modeling over-, equi- and underdispersion, hence we calculate δ and μ^* for various values of λ and θ. Table 6.2 summarizes the results obtained. It can be seen that, analogously to the 1-displaced GP model, the size-biased competitor (6.14) possesses the property of overdispersion for $0 < \lambda < 1$, whereas when $\lambda = 0$ we have Poisson limit. However, for negative values of λ apart from underdispersion we obtain particular areas where $\delta > 1$ and $\delta = 0$ (cf. violet marked cases in Table 6.2). Figures 6.5 and 6.6 locate these areas for the parameter range relevant for our further research by visualizing all parameter pairs for which $0 < \delta < 1$ as red areas, and all those with $\delta = 0$ as white areas. Notice that $\delta = 0$ corresponds to the case of one-point distribution, having variance equal to zero. Different blue toned areas point up the cases where $\delta > 1$, whereas the darkest ones signify cases with $\delta \geq 4$.

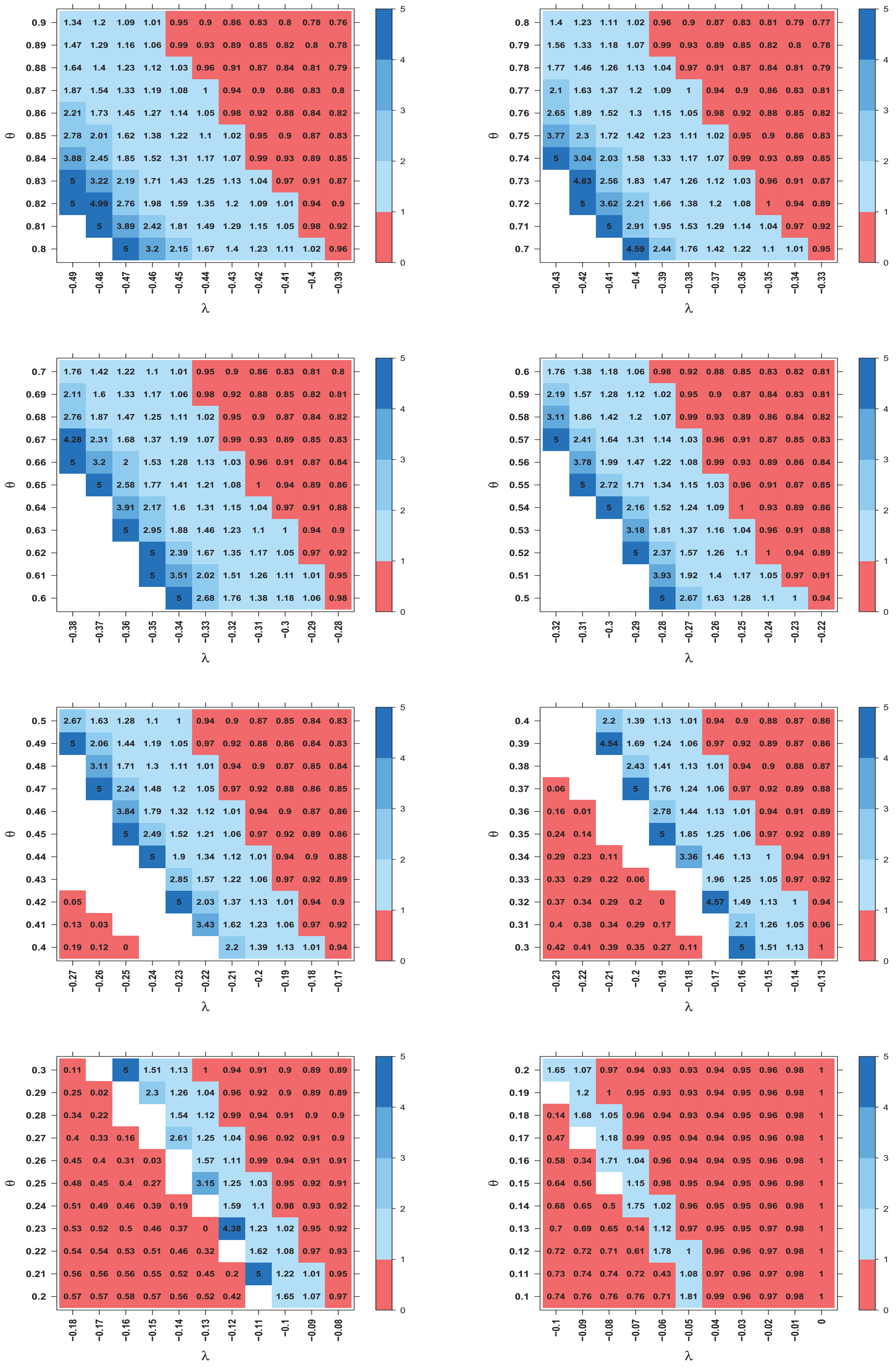
Table 6.2: Under- and overdispersion in the size-biased generalized Poisson distribution

					θ					
λ		0.1	0.5	0.7	0.82	0.98	1	1.01	1.6	1.8
-0.4	μ^*	0.58	0.87	1.01	1.10	1.21	1.22	1.23	1.65	1.80
	δ	0.41	0.20	**4.59**	0.94	0.71	0.70	0.69	0.57	0.56
-0.3	μ^*	0.67	0.98	1.13	1.22	1.35	1.36	1.37	1.82	1.98
	δ	0.50	**0.00**	0.83	0.73	0.68	0.68	0.68	0.63	0.62
-0.15	μ^*	0.84	1.19	1.36	1.47	1.61	1.63	**1.63**	2.15	2.32
	δ	0.67	0.82	0.79	0.78	0.78	0.78	**0.78**	0.77	0.77
-0.01	μ^*	1.08	1.48	1.67	1.79	1.95	1.97	1.98	2.56	2.76
	δ	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
0	μ^*	1.10	1.50	1.70	1.82	1.98	2.00	2.01	2.60	2.80
	δ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	μ^*	1.12	1.53	1.73	**1.85**	2.01	2.03	2.04	2.64	2.84
	δ	1.02	1.02	1.02	**1.02**	1.02	1.02	1.02	1.02	1.02
0.13	μ^*	1.44	1.90	2.13	2.26	**2.45**	2.47	2.48	3.16	3.39
	δ	1.39	1.35	1.35	1.34	**1.34**	1.34	1.34	1.33	1.33
0.2	μ^*	1.69	2.19	2.44	2.59	2.79	2.81	2.82	3.56	3.81
	δ	1.70	1.64	1.63	1.62	1.62	1.62	1.62	1.60	1.60
0.5	μ^*	4.20	5.00	5.40	5.64	5.96	6.00	6.02	7.20	7.60
	δ	5.25	5.00	4.91	4.86	4.81	4.80	4.80	4.65	4.61

The individual size-biased GP probabilities are dependent on previous occurrences and thus can be calculated using the ratio of the successive probabilities given by

$$\frac{\pi^*_{x|\lambda,\theta}}{\pi^*_{x-1|\lambda,\theta}} = \frac{e^{-\lambda}(\theta + x\lambda)^{x-1}}{(x-1)(\theta + x\lambda - \lambda)^{x-2}}\,, \quad x \geq 2\,, \tag{6.19}$$

where $\pi^*_{1|\lambda,\theta} = (1-\lambda)e^{-(\theta+\lambda)}$ is given by (6.14).



θ	−0.49	−0.48	−0.47	−0.46	−0.45	−0.44	−0.43	−0.42	−0.41	−0.4	−0.39
0.9	1.34	1.2	1.09	1.01	0.95	0.9	0.86	0.83	0.8	0.78	0.76
0.89	1.47	1.29	1.16	1.06	0.99	0.93	0.89	0.85	0.82	0.8	0.78
0.88	1.64	1.4	1.23	1.12	1.03	0.96	0.91	0.87	0.84	0.81	0.79
0.87	1.87	1.54	1.33	1.19	1.08	1	0.94	0.9	0.86	0.83	0.8
0.86	2.21	1.73	1.45	1.27	1.14	1.05	0.98	0.92	0.88	0.84	0.82
0.85	2.78	2.01	1.62	1.38	1.22	1.1	1.02	0.95	0.9	0.87	0.83
0.84	3.88	2.45	1.85	1.52	1.31	1.17	1.07	0.99	0.93	0.89	0.85
0.83	5	3.22	2.19	1.71	1.43	1.25	1.13	1.04	0.97	0.91	0.87
0.82	5	4.99	2.76	1.98	1.59	1.35	1.2	1.09	1.01	0.94	0.9
0.81		5	3.89	2.42	1.81	1.49	1.29	1.15	1.05	0.98	0.92
0.8			5	3.2	2.15	1.67	1.4	1.23	1.11	1.02	0.96

θ	−0.43	−0.42	−0.41	−0.4	−0.39	−0.38	−0.37	−0.36	−0.35	−0.34	−0.33
0.8	1.4	1.23	1.11	1.02	0.96	0.9	0.87	0.83	0.81	0.79	0.77
0.79	1.56	1.33	1.18	1.07	0.99	0.93	0.89	0.85	0.82	0.8	0.78
0.78	1.77	1.46	1.26	1.13	1.04	0.97	0.91	0.87	0.84	0.81	0.79
0.77	2.1	1.63	1.37	1.2	1.09	1	0.94	0.9	0.86	0.83	0.81
0.76	2.65	1.89	1.52	1.3	1.15	1.05	0.98	0.92	0.88	0.85	0.82
0.75	3.77	2.3	1.72	1.42	1.23	1.11	1.02	0.95	0.9	0.86	0.83
0.74	5	3.04	2.03	1.58	1.33	1.17	1.07	0.99	0.93	0.89	0.85
0.73		4.83	2.56	1.83	1.47	1.26	1.12	1.03	0.96	0.91	0.87
0.72		5	3.62	2.21	1.66	1.38	1.2	1.08	1	0.94	0.89
0.71			5	2.91	1.95	1.53	1.29	1.14	1.04	0.97	0.92
0.7				4.59	2.44	1.76	1.42	1.22	1.1	1.01	0.95

θ	−0.38	−0.37	−0.36	−0.35	−0.34	−0.33	−0.32	−0.31	−0.3	−0.29	−0.28
0.7	1.76	1.42	1.22	1.1	1.01	0.95	0.9	0.86	0.83	0.81	0.8
0.69	2.11	1.6	1.33	1.17	1.06	0.98	0.92	0.88	0.85	0.82	0.81
0.68	2.76	1.87	1.47	1.25	1.11	1.02	0.95	0.9	0.87	0.84	0.82
0.67	4.28	2.31	1.68	1.37	1.19	1.07	0.99	0.93	0.89	0.85	0.83
0.66	5	3.2	2	1.53	1.28	1.13	1.03	0.96	0.91	0.87	0.84
0.65		5	2.58	1.77	1.41	1.21	1.08	1	0.94	0.89	0.86
0.64			3.91	2.17	1.6	1.31	1.15	1.04	0.97	0.91	0.88
0.63			5	2.95	1.88	1.46	1.23	1.1	1	0.94	0.9
0.62				5	2.39	1.67	1.35	1.17	1.05	0.97	0.92
0.61				5	3.51	2.02	1.51	1.26	1.11	1.01	0.95
0.6					5	2.68	1.76	1.38	1.18	1.06	0.98

θ	−0.32	−0.31	−0.3	−0.29	−0.28	−0.27	−0.26	−0.25	−0.24	−0.23	−0.22
0.6	1.76	1.38	1.18	1.06	0.98	0.92	0.88	0.85	0.83	0.82	0.81
0.59	2.19	1.57	1.28	1.12	1.02	0.95	0.9	0.87	0.84	0.83	0.81
0.58	3.11	1.86	1.42	1.2	1.07	0.99	0.93	0.89	0.86	0.84	0.82
0.57	5	2.41	1.64	1.31	1.14	1.03	0.96	0.91	0.87	0.85	0.83
0.56		3.78	1.99	1.47	1.22	1.08	0.99	0.93	0.89	0.86	0.84
0.55		5	2.72	1.71	1.34	1.15	1.03	0.96	0.91	0.87	0.85
0.54			5	2.16	1.52	1.24	1.09	1	0.93	0.89	0.86
0.53				3.18	1.81	1.37	1.16	1.04	0.96	0.91	0.88
0.52				5	2.37	1.57	1.26	1.1	1	0.94	0.89
0.51					3.93	1.92	1.4	1.17	1.05	0.97	0.91
0.5					5	2.67	1.63	1.28	1.1	1	0.94

θ	−0.27	−0.26	−0.25	−0.24	−0.23	−0.22	−0.21	−0.2	−0.19	−0.18	−0.17
0.5	2.67	1.63	1.28	1.1	1	0.94	0.9	0.87	0.85	0.84	0.83
0.49	5	2.06	1.44	1.19	1.05	0.97	0.92	0.88	0.86	0.84	0.83
0.48		3.11	1.71	1.3	1.11	1.01	0.94	0.9	0.87	0.85	0.84
0.47		5	2.24	1.48	1.2	1.05	0.97	0.92	0.88	0.86	0.85
0.46			3.84	1.79	1.32	1.12	1.01	0.94	0.9	0.87	0.86
0.45			5	2.49	1.52	1.21	1.06	0.97	0.92	0.89	0.86
0.44				5	1.9	1.34	1.12	1.01	0.94	0.9	0.88
0.43					2.85	1.57	1.22	1.06	0.97	0.92	0.89
0.42	0.05				5	2.03	1.37	1.13	1.01	0.94	0.9
0.41	0.13	0.03				3.43	1.62	1.23	1.06	0.97	0.92
0.4	0.19	0.12	0				2.2	1.39	1.13	1.01	0.94

θ	−0.23	−0.22	−0.21	−0.2	−0.19	−0.18	−0.17	−0.16	−0.15	−0.14	−0.13
0.4			2.2	1.39	1.13	1.01	0.94	0.9	0.88	0.87	0.86
0.39			4.54	1.69	1.24	1.06	0.97	0.92	0.89	0.87	0.86
0.38				2.43	1.41	1.13	1.01	0.94	0.9	0.88	0.87
0.37	0.06			5	1.76	1.24	1.06	0.97	0.92	0.89	0.88
0.36	0.16	0.01			2.78	1.44	1.13	1.01	0.94	0.91	0.89
0.35	0.24	0.14			5	1.85	1.25	1.06	0.97	0.92	0.89
0.34	0.29	0.23	0.11			3.36	1.46	1.13	1	0.94	0.91
0.33	0.33	0.29	0.22	0.06			1.96	1.25	1.05	0.97	0.92
0.32	0.37	0.34	0.29	0.2	0		4.57	1.49	1.13	1	0.94
0.31	0.4	0.38	0.34	0.29	0.17			2.1	1.26	1.05	0.96
0.3	0.42	0.41	0.39	0.35	0.27	0.11		5	1.51	1.13	1

θ	−0.18	−0.17	−0.16	−0.15	−0.14	−0.13	−0.12	−0.11	−0.1	−0.09	−0.08
0.3	0.11		5	1.51	1.13	1	0.94	0.91	0.9	0.89	0.89
0.29	0.25	0.02		2.3	1.26	1.04	0.96	0.92	0.9	0.89	0.89
0.28	0.34	0.22			1.54	1.12	0.99	0.94	0.91	0.9	0.9
0.27	0.4	0.33	0.16		2.61	1.25	1.04	0.96	0.92	0.91	0.9
0.26	0.45	0.4	0.31	0.03		1.57	1.11	0.99	0.94	0.91	0.91
0.25	0.48	0.45	0.4	0.27		3.15	1.25	1.03	0.95	0.92	0.91
0.24	0.51	0.49	0.46	0.39	0.19		1.59	1.1	0.98	0.93	0.92
0.23	0.53	0.52	0.5	0.46	0.37	0	4.38	1.23	1.02	0.95	0.92
0.22	0.54	0.54	0.53	0.51	0.46	0.32		1.62	1.08	0.97	0.93
0.21	0.56	0.56	0.56	0.55	0.52	0.45	0.2	5	1.22	1.01	0.95
0.2	0.57	0.57	0.58	0.57	0.56	0.52	0.42		1.65	1.07	0.97

θ	−0.1	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0
0.2	1.65	1.07	0.97	0.94	0.93	0.93	0.94	0.95	0.96	0.98	1
0.19		1.2	1	0.95	0.93	0.93	0.94	0.95	0.96	0.98	1
0.18	0.14	1.68	1.05	0.96	0.94	0.93	0.94	0.95	0.96	0.98	1
0.17	0.47		1.18	0.99	0.95	0.94	0.94	0.95	0.96	0.98	1
0.16	0.58	0.34	1.71	1.04	0.96	0.94	0.94	0.95	0.96	0.98	1
0.15	0.64	0.56		1.15	0.98	0.95	0.94	0.95	0.96	0.98	1
0.14	0.68	0.65	0.5	1.75	1.02	0.96	0.95	0.95	0.96	0.98	1
0.13	0.7	0.69	0.65	0.14	1.12	0.97	0.95	0.95	0.97	0.98	1
0.12	0.72	0.72	0.71	0.61	1.78	1	0.96	0.96	0.97	0.98	1
0.11	0.73	0.74	0.74	0.72	0.43	1.08	0.97	0.96	0.97	0.98	1
0.1	0.74	0.76	0.76	0.76	0.71	1.81	0.99	0.96	0.97	0.98	1

Figure 6.5: Values of δ for the size-biased GP model when $\lambda \in [-0.49, 0]$, $\theta \in [0.1, 0.9]$

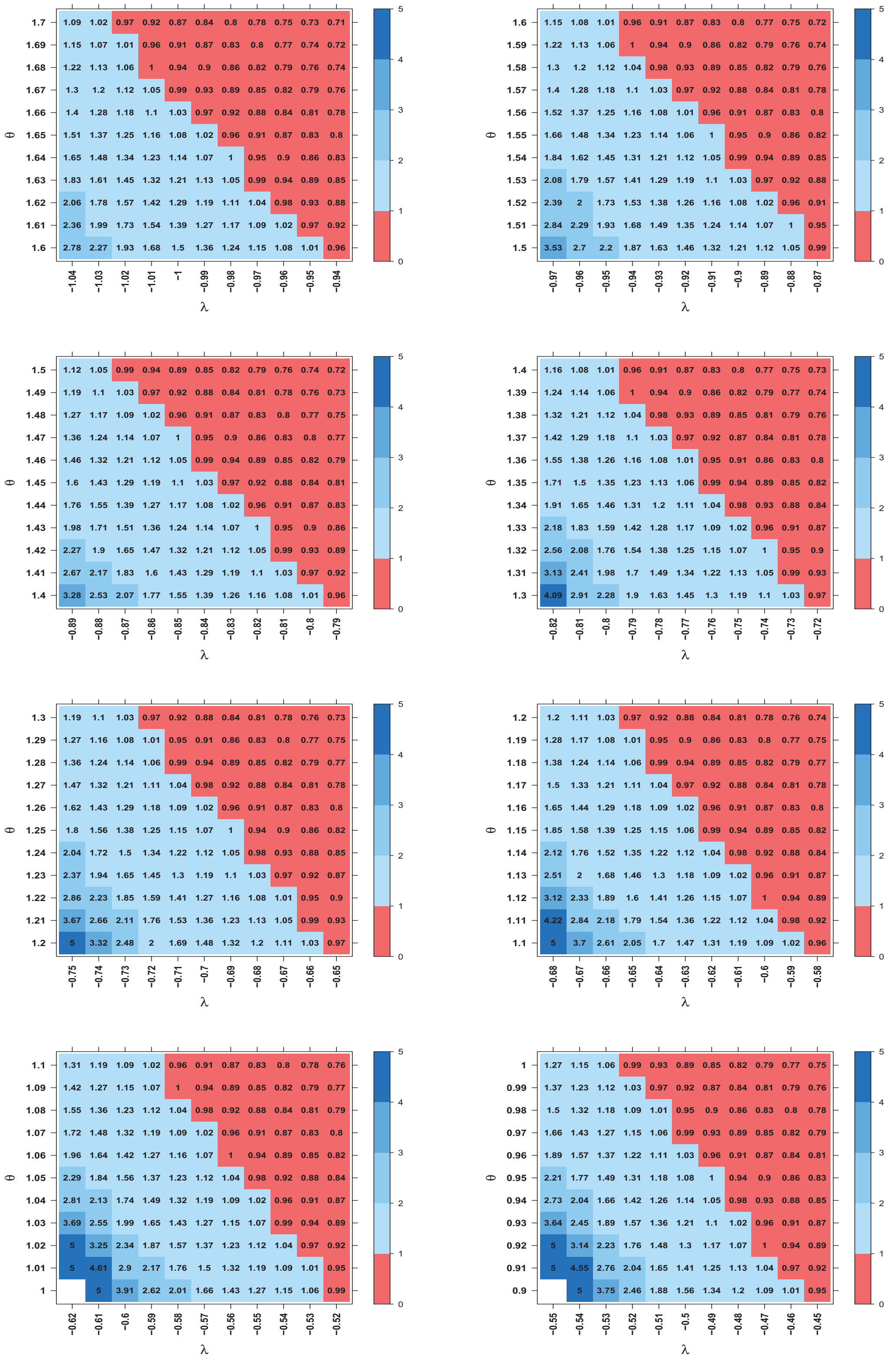


Figure 6.6: Values of δ for the size-biased GP model when $\lambda \in [-0.45, 1.04]$, $\theta \in [0.9, 1.7]$

6.4 Parameter Estimation

In this section we consider the three estimation procedures introduced in Section 2.4 and apply them to the 1-displaced and size-biased GP models.

6.4.1 Estimation by Method of Moments

The moment estimators of the unknown parameters λ and θ, for the *1-displaced GP model* (6.6), are obtained by equating the sample mean $\bar{x}$ and the sample variance m_2 to the corresponding theoretical counterparts defined by the first and second equation of (6.9), respectively. Solving such a system of equations, the moment estimators are obtained as

$$\hat{\theta}_{\text{MM}} = \sqrt{\frac{(\bar{x}-1)^3}{m_2}} \quad \text{and} \quad \hat{\lambda}_{\text{MM}} = 1 - \sqrt{\frac{\bar{x}-1}{m_2}}\,. \tag{6.20}$$

In the same manner we can determine the moment estimators for the *size-biased GP model*, defined in (6.14). In this case, the equating equations are given by

$$\bar{x}(1-\lambda)^2 = \theta(1-\lambda)+1\,, \quad m_2(1-\lambda)^4 = \theta(1-\lambda)+2\lambda\,. \tag{6.21}$$

Substituting $a = 1-\lambda$ in the equations above and eliminating parameter θ we obtain the following polynomial equation of the fourth degree

$$m_2 a^4 - \bar{x}a^2 + 2a - 1 = 0\,. \tag{6.22}$$

This equation can be solved easily in software `R` by executing a function `uniroot()`. Another possibility is to use e.g. the Newton-Raphson iterative method. It is obvious that the algebraic equation above has four solutions for a. Nevertheless, the required a has to be a positive real root of (6.22) such that $0 < a < \max(1, \theta/m) + 1$, as predetermined by the definition of the parameter λ. With a calculated in this way, the moment estimates of parameters λ and θ are given as

$$\hat{\lambda}_{\text{MM}} = 1 - a \quad \text{and} \quad \hat{\theta}_{\text{MM}} = \frac{\bar{x}a^2 - 1}{a}\,. \tag{6.23}$$

6.4.2 Estimation by Maximum Likelihood

Consider a random sample of size n from the *1-displaced GP model*, defined by (6.6). To find the maximum likelihood (ML) estimators of the parameters λ and θ we begin with the following likelihood function

$$L(\lambda, \theta | f_1, \ldots, f_k) = \prod_{i=1}^{k} \left(\frac{\theta\,(\theta + (i-1)\lambda)^{i-2}\, e^{-(\theta+(i-1)\lambda)}}{(i-1)!} \right)^{f_i}, \tag{6.24}$$

and maximize its log-likelihood function given by

$$l(\lambda,\theta|f_1,\ldots,f_k) = \log L(\lambda,\theta|f_1,\ldots,f_k) = \tag{6.25}$$

$$= n\log(\theta) + \sum_{i=1}^{k} f_i(i-2)\log\left(\theta+(i-1)\lambda\right) - \sum_{i=1}^{k} f_i\left(\theta+(i-1)\lambda\right) - \sum_{i=1}^{k} f_i \log\Gamma(i)\,,$$

where f_i are the observed frequencies such that $\sum_{i=1}^{k} f_i = n$ with k being the largest observation. We differentiate (6.25) partially with respect to the parameters λ and θ, and equate the resulting expressions to zero to obtain the ML equations as

$$\frac{\partial l(\lambda,\theta|f_1,\ldots,f_k)}{\partial\lambda} = \sum_{i=1}^{k}\frac{f_i(i-2)(i-1)}{\theta+(i-1)\lambda} - n(\bar{x}-1) = 0\,, \tag{6.26}$$

$$\frac{\partial l(\lambda,\theta|f_1,\ldots,f_k)}{\partial\theta} = \frac{n}{\theta} + \sum_{i=1}^{k}\frac{f_i(i-2)}{\theta+(i-1)\lambda} - n = 0\,. \tag{6.27}$$

The second-order partial derivatives of the log-likelihood function are given by

$$\frac{\partial^2 l(\lambda,\theta|\mathbf{f})}{\partial\lambda^2} = -\sum_{i=1}^{k}\frac{f_i(i-2)(i-1)^2}{\left(\theta+(i-1)\lambda\right)^2}\,, \quad \frac{\partial^2 l(\lambda,\theta|\mathbf{f})}{\partial\lambda\partial\theta} = -\sum_{i=1}^{k}\frac{f_i(i-2)(i-1)}{\left(\theta+(i-1)\lambda\right)^2}\,,$$

$$\frac{\partial^2 l(\lambda,\theta|\mathbf{f})}{\partial\theta^2} = -\frac{n}{\theta^2} - \sum_{i=1}^{k}\frac{f_i(i-2)}{\left(\theta+(i-1)\lambda\right)^2}\,, \tag{6.28}$$

where $\mathbf{f} = (f_1,\ldots,f_k)$, and obviously all of them are negative when $k>2$. Therefore, all the points (λ,θ) being solutions of the ML equations above are maxima of the log-likelihood function. Multiplying (6.26) by λ and (6.27) by θ, as suggested by Consul (1989, p. 102) for the unmodified distribution, and adding them afterwards results in the simple ML estimator of θ as

$$\hat{\theta}_{\text{ML}} = (\bar{x}-1)(1-\lambda)\,, \tag{6.29}$$

which when substituted in the equation (6.26) yields the ML estimate $\hat{\lambda}_{\text{ML}}$ as a solution of the following equation in λ

$$H(\lambda) = \sum_{i=1}^{k}\frac{f_i(i-2)(i-1)}{(\bar{x}-1)+(i-\bar{x})\lambda} - n(\bar{x}-1) = 0\,. \tag{6.30}$$

It can be seen that for $k\in\{1,2\}$ the sum above will become zero, and hence the ML estimates $\hat{\lambda}_{\text{ML}}$ and $\hat{\theta}_{\text{ML}}$ will not exist in this case. However, when the observed data sample has at least three non-zero frequency classes (i.e. $k>2$), the equation (6.30) has a unique root of λ. To prove this statement notice that for $\lambda\in\{0,1\}$ we have

$$H(1) = \sum_{i=1}^{k} f_i(i-2) - n(\bar{x}-1) = -n < 0\,, \tag{6.31}$$

$$H(0) = \frac{1}{\bar{x}-1}\left(\sum_{i=1}^{k} f_i(i-2)(i-1) - n(\bar{x}-1)^2\right) = n\left(\frac{m_2}{\bar{x}-1} - 1\right), \tag{6.32}$$

hence iff $d = m_2/(\bar{x} - 1) \geq 1$ (i.e. equi- or overdispersed data case) there must be a single real value of λ in $[0, 1)$ such that $H(\lambda) = 0$ (cf. Consul and Shoukri, 1984). To obtain such a unique value of λ the equation (6.30) has to be solved iteratively by applying the Newton-Raphson algorithm. As starting value of the iteration process one can use e.g. the moment estimator $\hat{\lambda}_{\text{MM}}$ given in (6.20). The value $\lambda^{(m)}$ at which the iteration process stops after m steps is then the required estimate $\hat{\lambda}_{\text{ML}}$, whereas $\hat{\theta}_{\text{ML}}$ is subsequently obtained from (6.29).

Furthermore, Consul and Famoye (1988) showed that the unique and admissible ML estimators $\hat{\lambda}_{\text{ML}}$ and $\hat{\theta}_{\text{ML}}$ also exist in the case when $\theta > 0$ and $\lambda < 0$, and that these can be obtained by using the same ML equations as in the case $0 \leq \lambda < 1$.

The ML estimates of λ and θ for the *size-biased GP model* (6.14), are obtained by maximizing the corresponding size-biased log-likelihood function defined by

$$l(\lambda, \theta|\mathbf{f}) = n \log(1-\lambda) + \sum_{i=1}^{k} f_i (i-1) \log(\theta + i\lambda) - n\theta - n\bar{x}\lambda - \sum_{i=1}^{k} f_i \log \Gamma(i) \,. \quad (6.33)$$

where $\mathbf{f} = (f_1, \ldots, f_k)$. By differentiating (6.33), the ML equations are obtained as

$$\frac{\partial l(\lambda, \theta|\mathbf{f})}{\partial \lambda} = \frac{n}{\lambda - 1} + \sum_{i=1}^{k} \frac{f_i (i-1) i}{\theta + i\lambda} - n\bar{x} = 0 \,, \quad (6.34)$$

$$\frac{\partial l(\lambda, \theta|\mathbf{f})}{\partial \theta} = \sum_{i=1}^{k} \frac{f_i (i-1)}{\theta + i\lambda} - n = 0 \,, \quad (6.35)$$

and the Newton-Raphson algorithm may be used to find the ML estimates $\hat{\lambda}_{\text{ML}}$ and $\hat{\theta}_{\text{ML}}$ as the simultaneous solutions of these two equations. Consequently, the second-order partial derivatives of the log-likelihood (6.33) required for this iterative process result in

$$\begin{aligned} \frac{\partial^2 l(\lambda, \theta|\mathbf{f})}{\partial \lambda^2} &= -\frac{n}{(\lambda - 1)^2} - \sum_{i=1}^{k} \frac{f_i (i-1) i^2}{(\theta + i\lambda)^2} \,, \\ \frac{\partial^2 l(\lambda, \theta|\mathbf{f})}{\partial \lambda \partial \theta} &= -\sum_{i=1}^{k} \frac{f_i (i-1) i}{(\theta + i\lambda)^2} \,, \quad \frac{\partial^2 l(\lambda, \theta|\mathbf{f})}{\partial \theta^2} = -\sum_{i=1}^{k} \frac{f_i (i-1)}{(\theta + i\lambda)^2} \,. \end{aligned} \quad (6.36)$$

However, the log-likelihood above can be also optimized with a numerical optimizer.

6.4.3 Estimation Based on Mean and First Frequency Class

To derive the parameter estimators based on mean and first frequency class for the *1-displaced GP model* (6.6) we equate the sample mean $\bar{x}$ and the relative frequency of the first class f_1/n to the theoretical mean $\mu^d = \theta/(1-\lambda) + 1$ and the probability

$\pi^d_{1|\lambda,\theta} = e^{-\theta}$ of the first class, respectively. Solving such a system of simultaneous equations we obtain very simple estimators of the parameters λ and θ as follows

$$\hat{\theta}_{\text{FF}} = \log\left(\frac{n}{f_1}\right) \quad \text{and} \quad \hat{\lambda}_{\text{FF}} = 1 - \frac{1}{\bar{x}-1}\log\left(\frac{n}{f_1}\right). \tag{6.37}$$

The explicit versions for the estimators $\hat{\lambda}_{\text{FF}}$ and $\hat{\theta}_{\text{FF}}$ are not available for the *size-biased GP model* (6.14), since they are solutions of the following system of equations

$$\bar{x} = \frac{\theta(1-\lambda)+1}{(1-\lambda)^2}, \quad \frac{f_1}{n} = (1-\lambda)e^{-\theta-\lambda}, \tag{6.38}$$

which after substituting the expression for θ given by the second equation above into the first equation results in the following complex transcendental equation in λ

$$(\bar{x}-1)\lambda^2 - (2\bar{x}+\log(f_1/n)-1)\lambda - (1-\lambda)\log(1-\lambda) + \bar{x} + \log(f_1/n) - 1 = 0\,. \tag{6.39}$$

This equation can be solved for $\hat{\lambda}_{\text{FF}}$ in the same manner as suggested for solving equation (4.19) from Chapter 4 to obtain $\hat{\theta}_{\text{ML}}$. However, we can use also the Newton-Raphson method instead. Knowing $\hat{\lambda}_{\text{FF}}$, the estimator $\hat{\theta}_{\text{FF}}$ is given by substituting it into one of the two equations in (6.38).

6.5 A Simulation Study

We investigate the accuracy of the previously mentioned estimation techniques for the 1-displaced and size-biased GP models by performing a simulation study where all three dispersion situations are considered. Since the comparison of the two distributions is reasonable only when some common characteristics are fixed, and both distributions are two-parametric, we fix the degree of dispersion δ and the first theoretical moment μ. If the true parameter values, specified in terms of δ and μ, are known, then we can easily find out the consequence of applying the wrong model and how effective the proposed estimation techniques are. With μ and δ known, the parameters of the 1-displaced GP model arise from the relations (6.9) and (6.10) as $\lambda = 1 - \sqrt{1/\delta}$ and $\theta = (\mu - 1)\sqrt{1/\delta}$. The model settings are given in the headings of Table 6.3. However, it is impossible to obtain explicit expressions for the theoretical parameter values of the size-biased GP model, because relation (6.18) for δ is considerably complex. Therefore, we make an exhaustive grid search through the parameter space to identify the proper (λ, θ) combinations for a particular pairs $(\mu,\ \delta)$, specified in advance. These model settings are obtained from Table 6.2 as green marked cases and displayed once again in the headings of Table 6.4. For each dispersion case $M = 500$ Monte Carlo samples of size $n = 500$ and $n = 1000$ are constructed from both models above and compared. To generate GP random variables we use the inversion method (cf. Stadlober, 1989), where the single probabilities are computed through the recurrence relationship defined above (see relations (6.13) and (6.19)). The mean values of $M = 500$ parameter estimates, denoted by $\bar{\lambda}$ and $\bar{\theta}$, as well as the estimated standard errors, denoted by $\mathrm{se}_{\bar{\lambda}}$ and $\mathrm{se}_{\bar{\theta}}$, have been calculated for each of the data situations considered here. Table 6.3 summarizes the simulation results for the 1-displaced GP model. We encountered no difficulty when performing these simulation experiments. All three estimation methods yield equally good results for both sample sizes, differing only in the third or fourth decimal place. Both parameters are estimated without bias. Nevertheless, the relative standard errors of $\bar{\theta}$ are much smaller than those of $\bar{\lambda}$, regardless of the estimation method and the dispersion case. The results for the size-biased GP model are presented in Table 6.4. It can be seen that the results obtained for the mean of 500 estimates $\hat{\lambda}$ (i.e. $\bar{\lambda}$) are consistent with those of the 1-displaced model. However, the relative

Table 6.3: Estimation results for data simulated from the 1-displaced GP model

$\delta = 1.34$	$(\lambda, \theta) = (0.14, 1.25)$					
$\mu = 2.45$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.137 (0.031)	1.256 (0.066)	0.137 (0.031)	1.256 (0.065)	0.136 (0.038)	1.256 (0.070)
$n = 1000$	0.140 (0.022)	1.251 (0.044)	0.140 (0.022)	1.251 (0.040)	0.140 (0.027)	1.251 (0.049)
$\delta = 1.02$	$(\lambda, \theta) = (0.01, 0.84)$					
$\mu = 1.85$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.009 (0.031)	0.842 (0.049)	0.008 (0.032)	0.842 (0.049)	0.009 (0.036)	0.841 (0.050)
$n = 1000$	0.011 (0.022)	0.840 (0.033)	0.010 (0.022)	0.840 (0.033)	0.011 (0.026)	0.839 (0.036)
$\delta = 0.78$	$(\lambda, \theta) = (-0.13, 0.71)$					
$\mu = 1.63$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	-0.130 (0.032)	0.711 (0.045)	-0.132 (0.031)	0.712 (0.045)	-0.130 (0.036)	0.711 (0.046)
$n = 1000$	-0.130 (0.023)	0.710 (0.031)	-0.131 (0.022)	0.711 (0.030)	-0.130 (0.026)	0.710 (0.032)

Table 6.4: Estimation results for data simulated from the size-biased GP model

$\delta = 1.34$	$(\lambda, \theta) = (0.13, 0.98)$					
$\mu = 2.45$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.127 (0.028)	0.991 (0.113)	0.127 (0.028)	0.993 (0.111)	0.126 (0.034)	0.994 (0.130)
$n = 1000$	0.130 (0.020)	0.981 (0.078)	0.130 (0.019)	0.981 (0.077)	0.129 (0.024)	0.982 (0.091)
$\delta = 1.02$	$(\lambda, \theta) = (0.01, 0.82)$					
$\mu = 1.85$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.008 (0.031)	0.825 (0.099)	0.007 (0.031)	0.827 (0.100)	0.009 (0.036)	0.823 (0.110)
$n = 1000$	0.010 (0.022)	0.819 (0.069)	0.010 (0.022)	0.820 (0.070)	0.010 (0.026)	0.818 (0.081)
$\delta = 0.78$	$(\lambda, \theta) = (-0.15, 1.01)$					
$\mu = 1.63$	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\lambda}_{\mathrm{FF}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	-0.153 (0.037)	1.019 (0.102)	-0.156 (0.036)	1.024 (0.099)	-0.154 (0.042)	1.020 (0.112)
$n = 1000$	-0.151 (0.026)	1.013 (0.071)	-0.153 (0.025)	1.016 (0.069)	-0.152 (0.030)	1.014 (0.079)

standard errors $\mathrm{RSE}_{\bar{\theta}}$ manifest rather increasing tendency compared to those of the 1-displaced case.

Attempting to evaluate whether the GP model or the SP model performs better under particular requirements, we fix again δ and μ to obtain the parametrization of the SP model as $\theta = \delta + \mu - 2$ and $\alpha = (\mu - 1)/\theta$ (see relations (4.4) and (4.7), Section 4.2). With these expressions, the SP model settings for δ and μ determined a priori are given in the headings of Table 6.5 showing simulation results obtained. It become apparent that the parameter estimates of θ are of similar precision as those obtained in Table 6.3 for the GP model in the over- and equidispersed data case, but get more imprecise as soon as the underdispersed data case is at hand. As to the RSE of the parameter α we found out that they are more precise compared to those of the parameter λ. For $\delta > 1$ we obtain 2 - 4% accuracy, for $\delta \approx 1$ approximately 4 - 7%, whereas for $\delta < 1$ around 8 - 12% estimation precision for α.

Table 6.5: Estimation results for data simulated from the 1-displaced SP model

$\delta = 1.34$	$(\alpha, \theta) = (0.81, 1.79)$			
$\mu = 2.45$	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$
$n = 500$	0.813 (0.034)	1.787 (0.093)	0.811 (0.027)	1.790 (0.080)
$n = 1000$	0.812 (0.024)	1.788 (0.067)	0.811 (0.019)	1.789 (0.057)
$\delta = 1.02$	$(\alpha, \theta) = (0.98, 0.87)$			
$\mu = 1.85$	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$
$n = 500$	0.989 (0.071)	0.867 (0.076)	0.985 (0.061)	0.869 (0.069)
$n = 1000$	0.984 (0.049)	0.869 (0.052)	0.982 (0.043)	0.870 (0.048)
$\delta = 0.78$	$(\alpha, \theta) = (1.54, 0.41)$			
$\mu = 1.63$	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$
$n = 500$	1.572 (0.192)	0.409 (0.054)	1.565 (0.179)	0.410 (0.052)
$n = 1000$	1.554 (0.133)	0.410 (0.039)	1.551 (0.124)	0.410 (0.037)

Figures 6.7, 6.8 and 6.9 contain nine plots showing estimation results corresponding to the model settings given in Table 6.3 for over-, equi- and underdispersed situations, respectively. In each figure the first two rows illustrate the distribution of the $M = 500$ parameter estimates $\hat{\lambda}$ and $\hat{\theta}$ for each of the estimation procedures applied. The third row demonstrates the corresponding scatterplots. The vertical solid black line pictured in the histograms shows the true parameter value, and the dashed black line displays the mean of the obtained estimates, i.e. $\bar{\lambda}$ and $\bar{\theta}$. In all cases these two lines coincide. Additionally, we plotted normal densities with the mean values $\bar{\lambda}$ and $\bar{\theta}$ as centers, whereas the standard deviation is obtained from the simulation variance of the estimates. In case of the ML estimation we plot also a second density, with the same center but standard deviation resulting from the variance-covariance matrix obtained by the Hessian. Notice that the differences between these two curves are negligible. The asymptotic normality of the estimates is given in all cases.

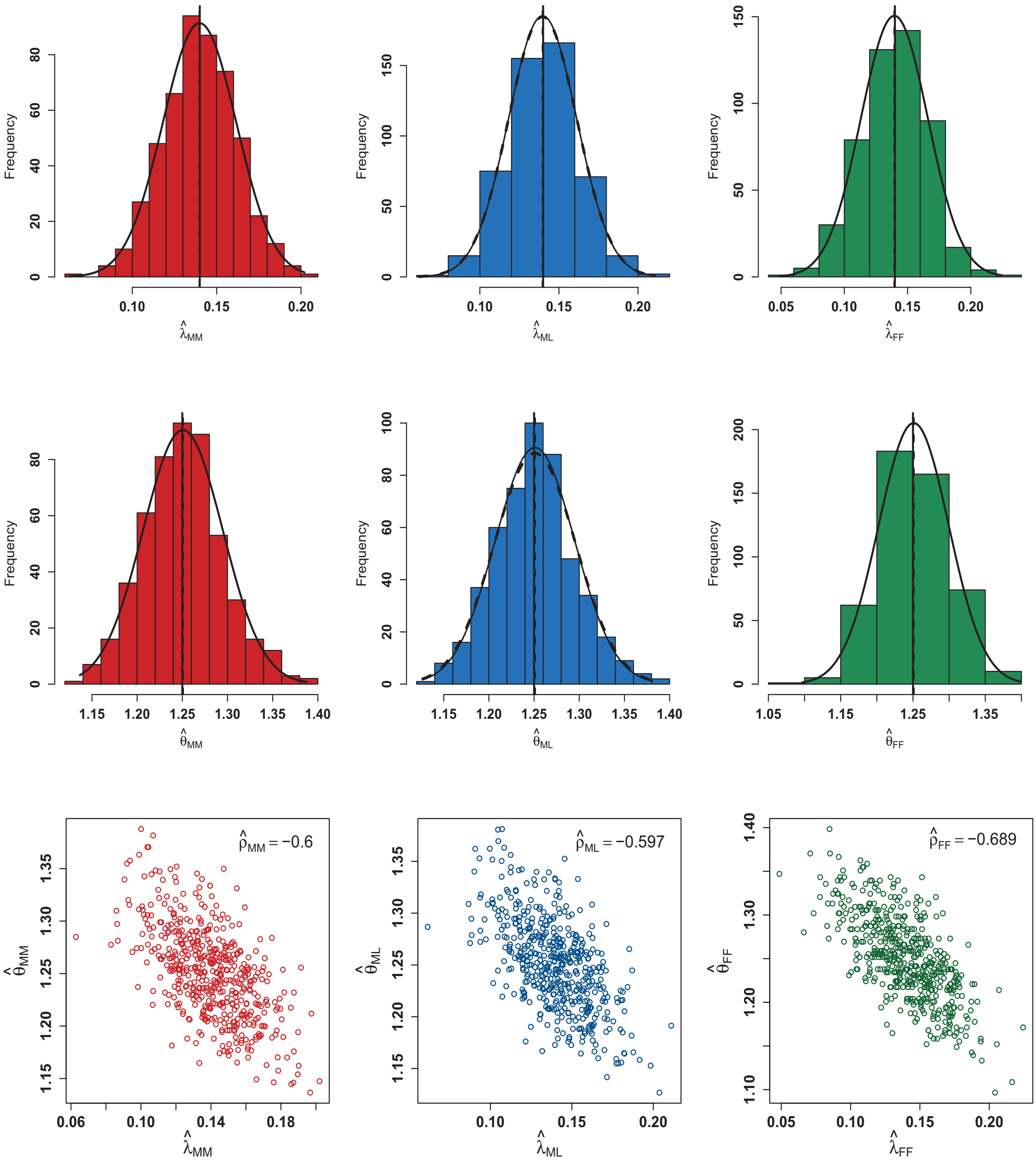


Figure 6.7: Estimation results obtained from $M = 500$ simulated 1-displaced GP samples of size $n = 1000$ with model parameters $(\lambda, \theta) = (0.14, 1.25)$.

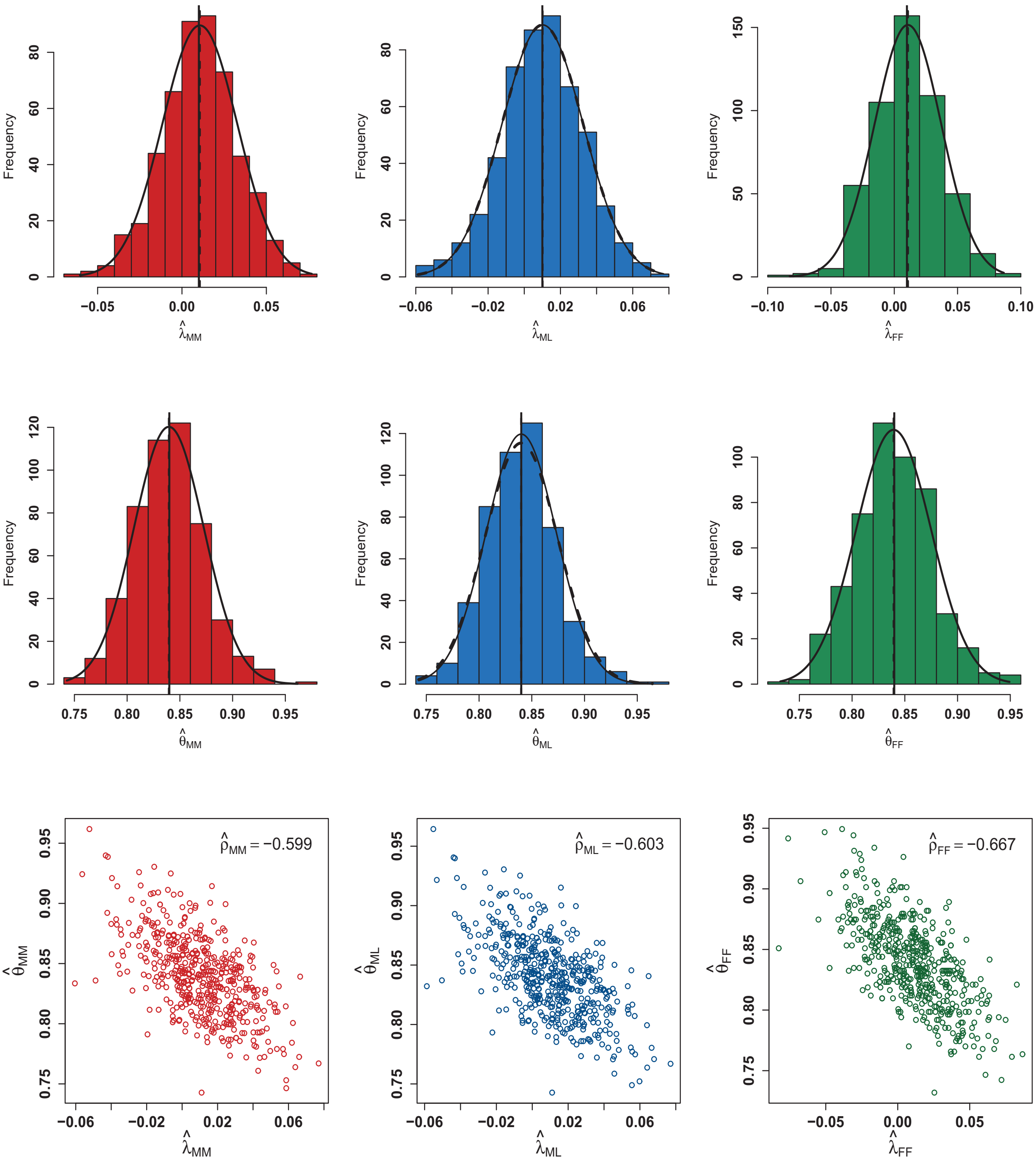


Figure 6.8: Estimation results obtained from $M = 500$ simulated 1-displaced GP samples of size $n = 1000$ with model parameters $(\lambda, \theta) = (0.01, 0.84)$.

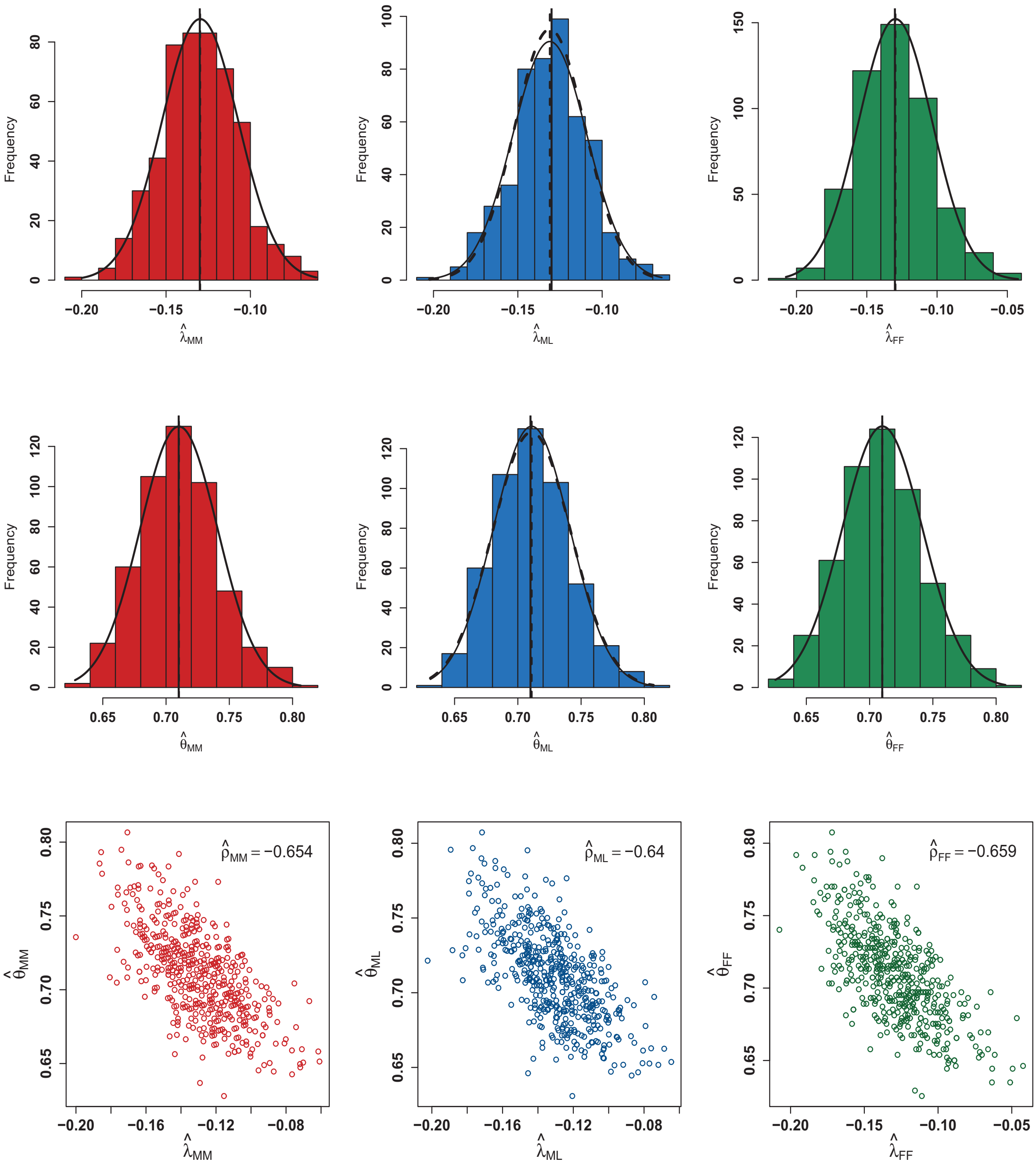


Figure 6.9: Estimation results obtained from $M = 500$ simulated 1-displaced GP samples of size $n = 1000$ with model parameters $(\lambda, \theta) = (-0.13, 0.71)$.

Figures 6.10, 6.11 and 6.12 visualize the simulation results corresponding to the settings given in Table 6.4 for over-, equi- and underdispersed situations, respectively. The scatterplots show very high correlation among the estimated parameter values for each of the estimation methods applied to all dispersion cases.

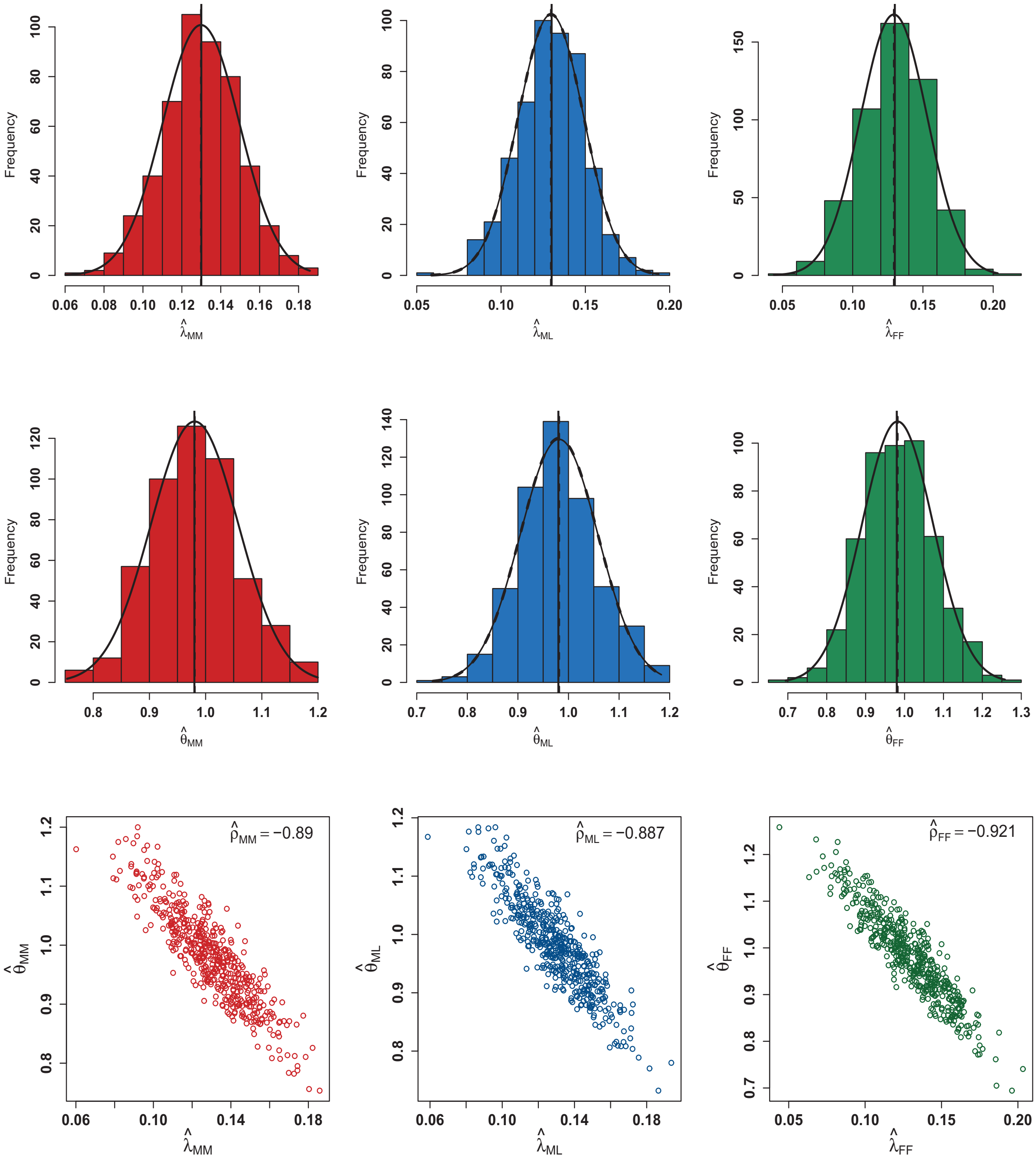


Figure 6.10: Estimation results obtained from $M = 500$ simulated size-biased GP samples of size $n = 1000$ with model parameters $(\lambda, \theta) = (0.13, 0.98)$.

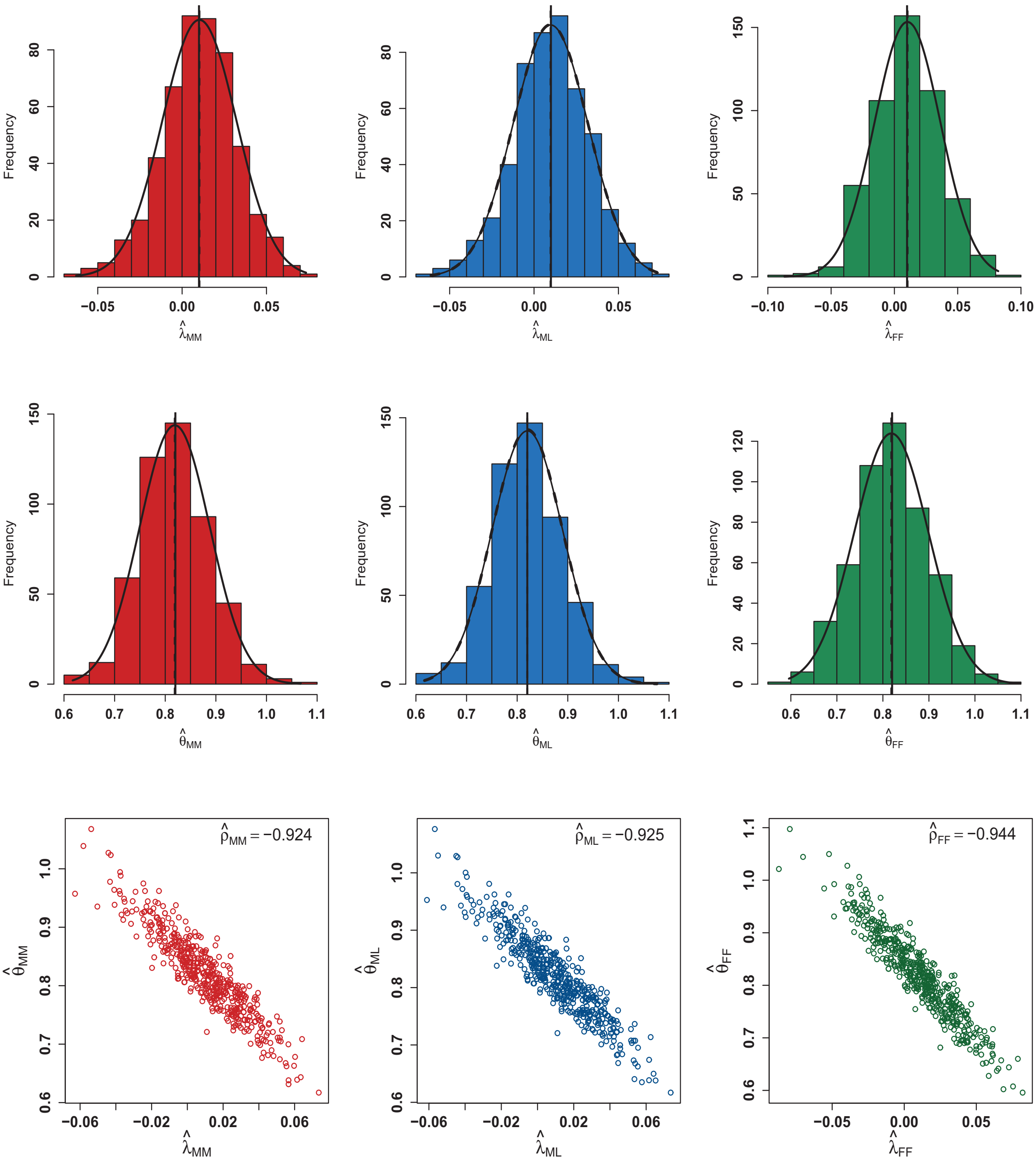


Figure 6.11: Estimation results obtained from $M = 500$ simulated size-biased GP samples of size $n = 1000$ with model parameters $(\lambda, \theta) = (0.01, 0.82)$.

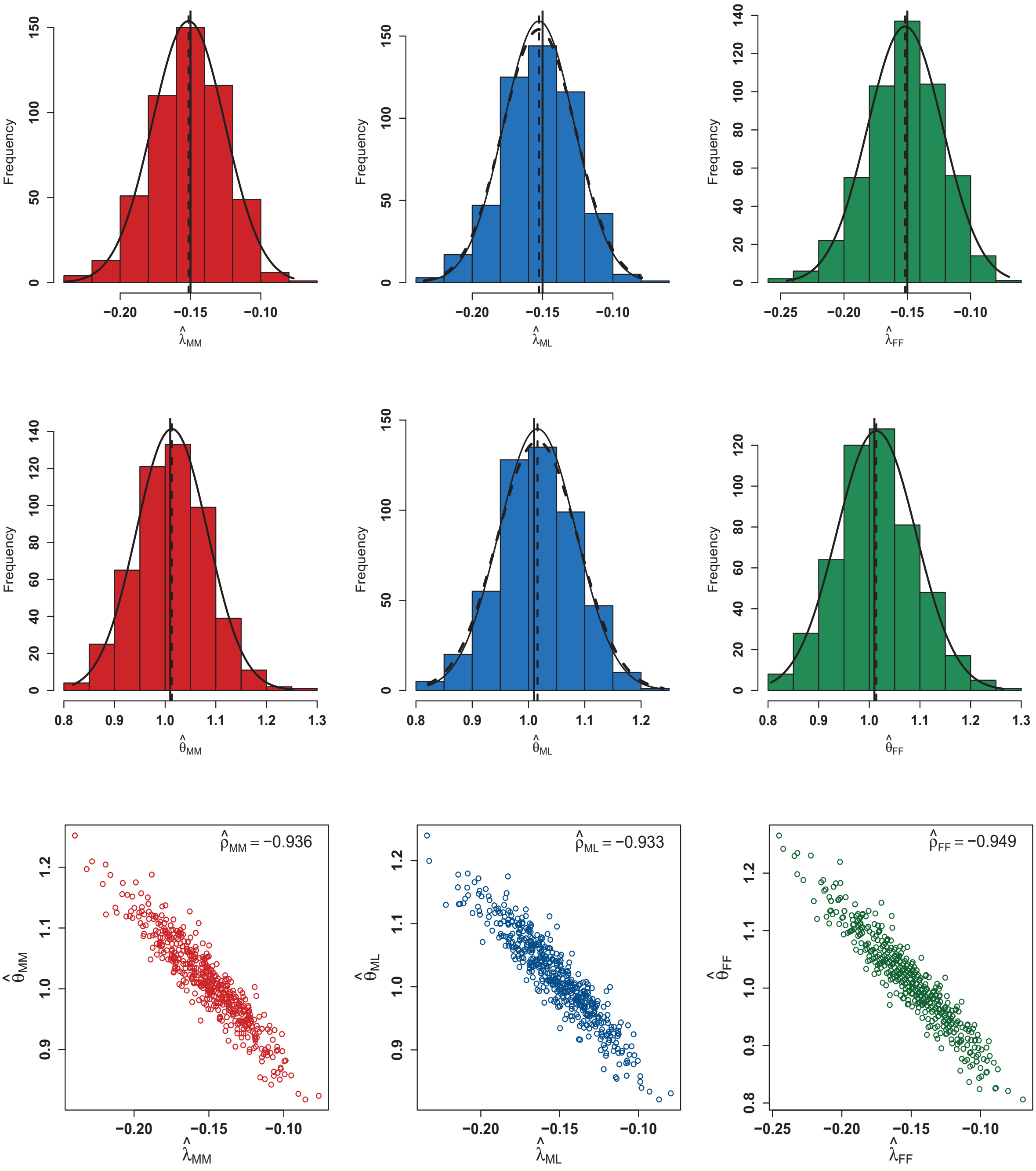


Figure 6.12: Estimation results obtained from $M = 500$ simulated size-biased GP samples of size $n = 1000$ with model parameters $(\lambda, \theta) = (-0.15, 1.01)$.

The estimation results for the 1-displaced SP model are graphically presented in Figures 6.13, 6.14, 6.15 and 6.16. Notice that in the case of $d < 1$ there is a nearly linear dependence between the estimated parameter values (cf. Figure 6.16).

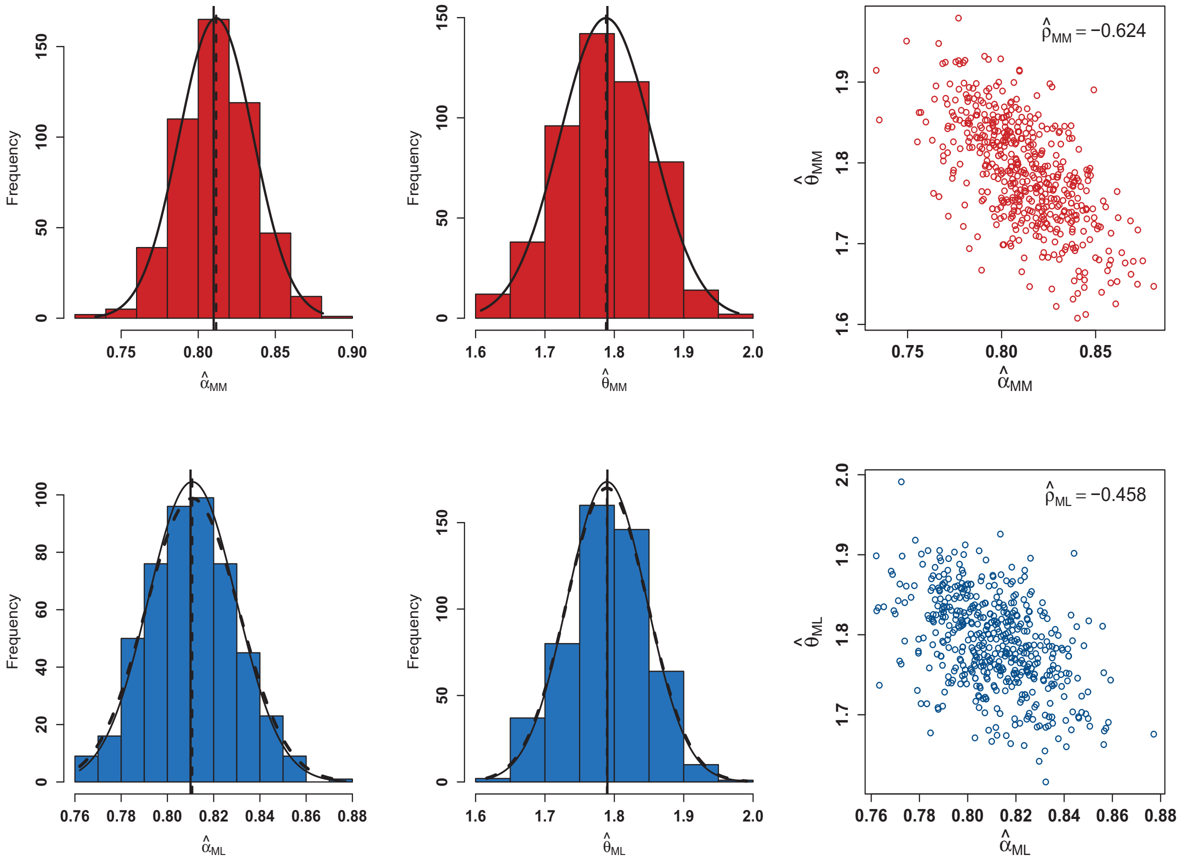


Figure 6.13: Estimation results obtained from $M = 500$ simulated 1-displaced SP samples of size $n = 1000$ with model parameters $(\alpha, \theta) = (0.81, 1.79)$.

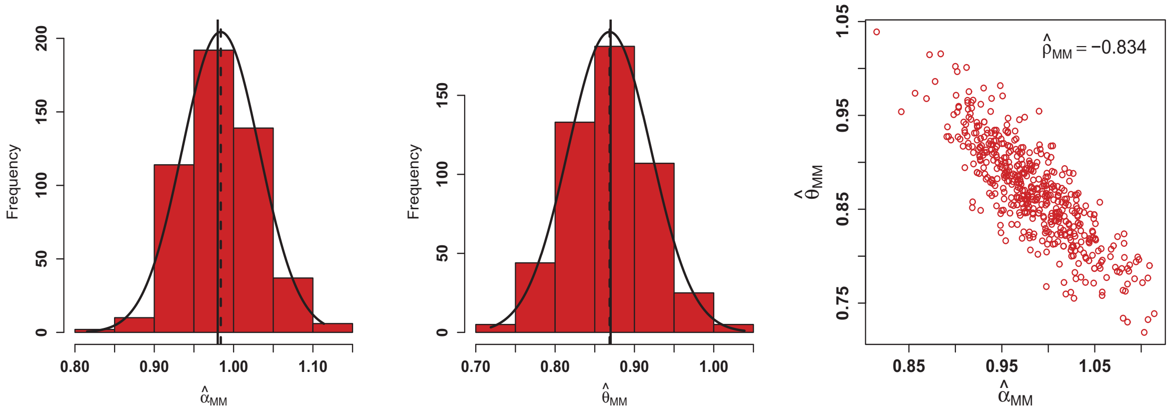


Figure 6.14: Estimation results obtained from $M = 500$ simulated 1-displaced SP samples of size $n = 1000$ with model parameters $(\alpha, \theta) = (0.98, 0.87)$.

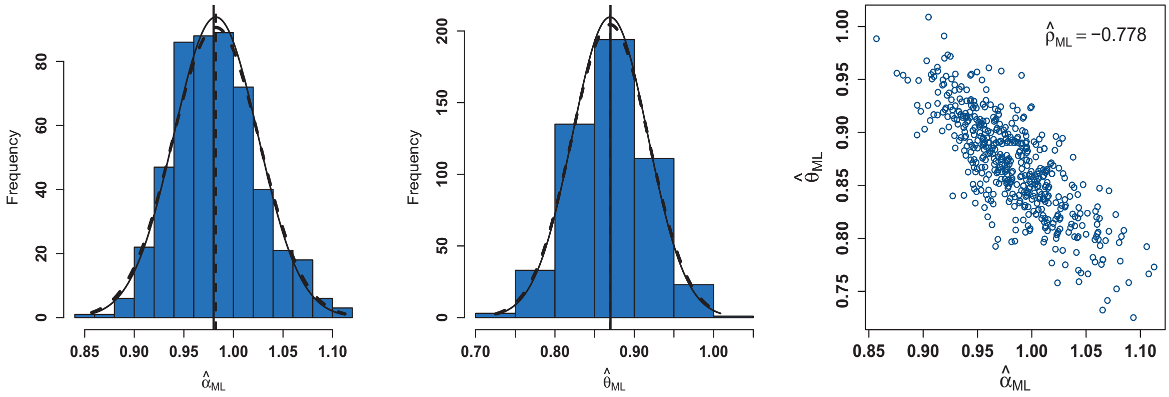


Figure 6.15: Estimation results obtained from $M = 500$ simulated 1-displaced SP samples of size $n = 1000$ with model parameters $(\alpha, \theta) = (0.98, 0.87)$.

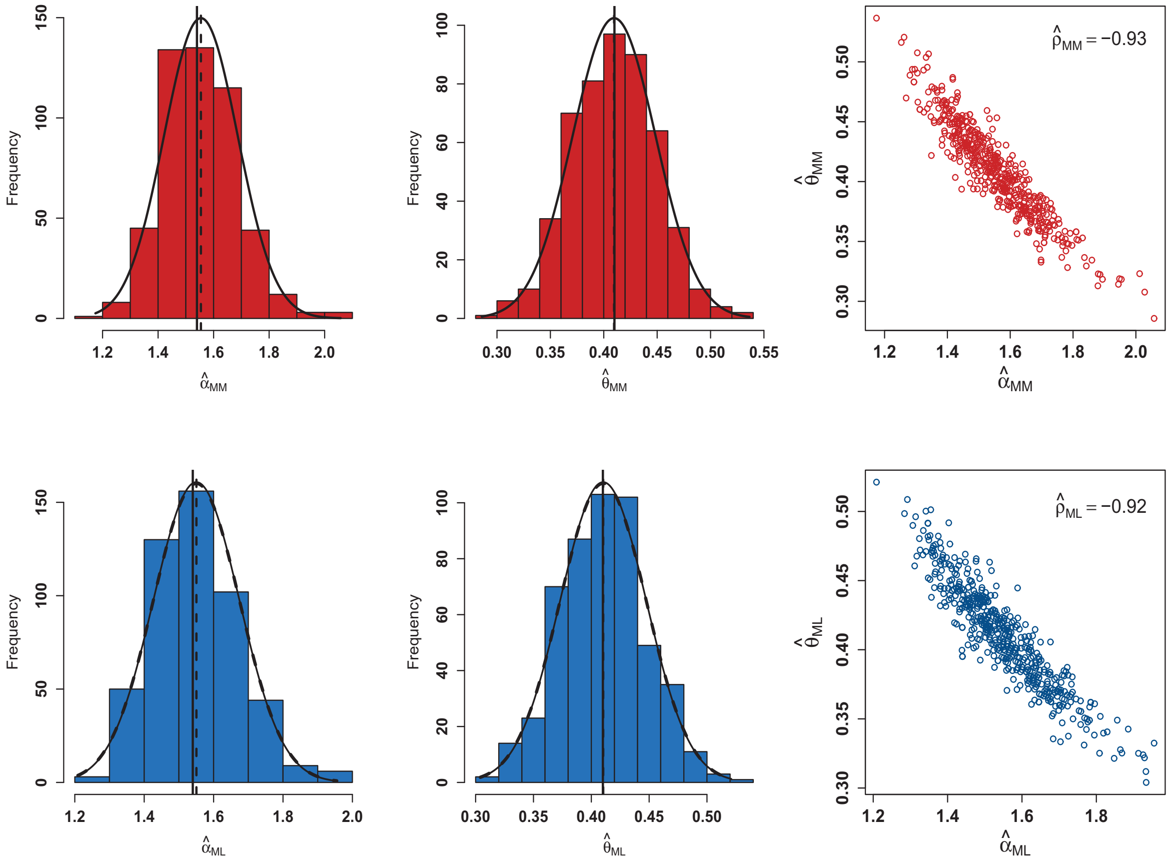


Figure 6.16: Estimation results obtained from $M = 500$ simulated 1-displaced SP samples of size $n = 1000$ with model parameters $(\alpha, \theta) = (1.54, 0.41)$.

Chapter 7

Cohen-Poisson Distribution

7.1 Introduction

The Cohen-Poisson (CP) distribution belongs to the class of misrecorded Poisson distributions. These altered distributions allow for errors in recording a variable that is in fact Poisson distributed. Specifically, the CP distribution has its application in situations where the true values of “ones” are erroneously recorded as “zeros” with probability α, while values of two and more are recorded correctly from a Poisson (cf. Johnson et al., 1992, p. 187). This distribution is a two-parametric modification of the Poisson distribution with nonnegative parameters, named here α and θ. The next sections provide evidence that its 1-displaced and size-biased versions are applicable to count data with overdispersion and underdispersion, respectively. Furthermore, for $\alpha = 0$ the CP distribution reduces to the standard Poisson distribution with parameter θ, the equidispersed case.

7.2 1-Displaced Cohen-Poisson Distribution

A discrete random variable X^d is said to have a 1-displaced CP distribution if its pmf is given by (cf. Wimmer and Altmann, 1999, p. 82)

$$\pi^d_{x|\alpha,\theta} = P(X^d = x) = \begin{cases} (1+\alpha\theta)e^{-\theta}, & x = 1\,, \\ (1-\alpha)\theta e^{-\theta}, & x = 2\,, \\ \theta^{x-1}e^{-\theta}/(x-1)!, & x = 3, 4, \dots\,, \end{cases} \tag{7.1}$$

where $\theta > 0$ and $0 \le \alpha \le 1$. Its pgf can be easily found by definition as

$$G_{X^d}(t) = \sum_{i=1}^{\infty} t^i \pi^d_{i|\alpha,\theta} = t(1+\alpha\theta)e^{-\theta} + t^2(1-\alpha)\theta e^{-\theta} + \sum_{i=3}^{\infty} t^i \frac{\theta^{i-1}e^{-\theta}}{(i-1)!}\,,$$

and since the above series expansion equals $t(e^{\theta} - t\theta e^{-\theta} - e^{-\theta})$, it becomes

$$G_{X^d}(t) = t\left(e^{\theta(t-1)} + \alpha\theta e^{-\theta}(1-t)\right)\,. \tag{7.2}$$

From this result, the mean and the variance are obtained as

$$\mu^d = \mathrm{E}(X^d) = 1 + \theta(1 - \alpha e^{-\theta}) \text{ and } \mathrm{var}(X^d) = \theta + \alpha\theta e^{-\theta}(2\theta - 1) - \alpha^2\theta^2 e^{-2\theta}\,. \quad (7.3)$$

It subsequently follows from the pgf that the factorial moments are given by

$$\mu^d_{(1)} = \mu^d\,,\ \mu^d_{(2)} = \theta^2 + 2\theta(1 - \alpha e^{-\theta})\,,\ \text{and } \mu^d_{(k)} = \theta^k + k\theta^{k-1}\,,\ \text{for } k \geq 3\,. \quad (7.4)$$

Furthermore, raw and central moments of distribution (7.1) can be easily found from the above factorial moments using relations (B.9) and (B.13), Appendix B. Computation of the successive probabilities is possible by applying the first three probabilities in (7.1) and the following recurrence relation for the remaining ones

$$\pi^d_{x|\alpha,\theta} = \frac{\theta}{x-1}\pi^d_{x-1|\alpha,\theta}\,,\ \text{for } x \geq 4\,. \quad (7.5)$$

This distribution has the ability to model overdispersion with respect to Poisson variation, which can be confirmed by the following index of dispersion

$$\delta = \frac{\mathrm{var}(X^d)}{\mathrm{E}(X^d) - 1} = 1 + \frac{\alpha\theta e^{-\theta}(2 - \alpha e^{-\theta})}{1 - \alpha e^{-\theta}} = 1 + g(\alpha, \theta). \quad (7.6)$$

Obviously, for $\alpha = 0$ we have $\delta = 1$ and $G_{X^d}(t) = te^{\theta(t-1)}$, hence distribution (7.1) simplifies to the 1-displaced Poisson distribution. Modifications in mean and index of dispersion for diverse choices of parameters α and θ are illustrated in Table 7.1. The left panel of Figure 7.1 clearly indicates that given a fixed value of θ, the value of the function $g(\alpha, \theta)$ increases, as α increases from 0 to 1. Consequently, the degree of overdispersion increases as well. However, for sufficiently large θ, the value of the

Table 7.1: 1-displaced Cohen-Poisson as adequate model choice for overdispersed data

		θ								
α		0.1	0.5	1	1.5	2	2.5	3	6	9
0	μ^d	1.10	1.50	2.00	2.50	3.00	3.50	4.00	7.00	10.00
	δ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	μ^d	1.08	1.44	1.93	2.43	2.95	3.46	3.97	7.00	10.00
	δ	1.04	1.13	1.15	1.14	1.11	1.08	1.06	1.01	1.00
0.4	μ^d	1.06	1.38	1.85	2.37	2.89	3.42	3.94	6.99	10.00
	δ	1.09	1.28	1.32	1.28	1.22	1.17	1.12	1.01	1.00
0.6	μ^d	1.05	1.32	1.78	2.30	2.84	3.38	3.91	6.99	10.00
	δ	1.17	1.47	1.50	1.43	1.34	1.25	1.18	1.02	1.00
0.8	μ^d	1.03	1.26	1.71	2.23	2.78	3.34	3.88	6.99	10.00
	δ	1.33	1.71	1.71	1.59	1.46	1.34	1.24	1.02	1.00
1	μ^d	1.01	1.20	1.63	2.17	2.73	3.29	3.85	6.99	10.00
	δ	2.04	2.07	1.95	1.77	1.58	1.43	1.31	1.03	1.00

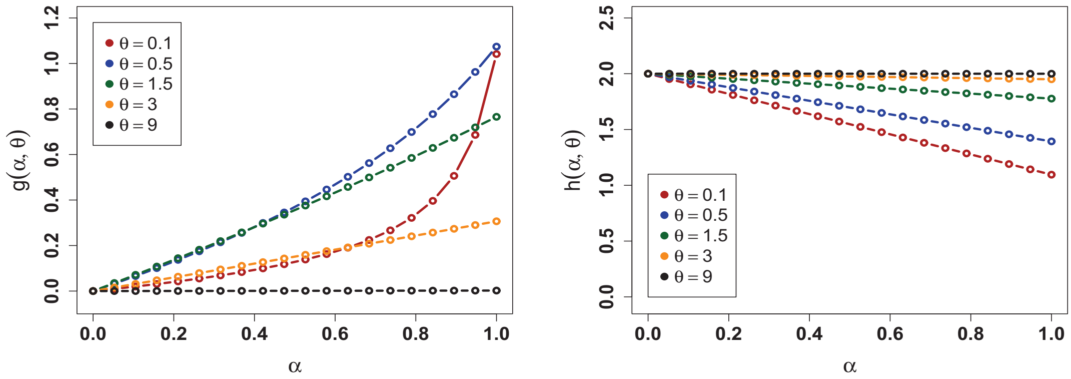


Figure 7.1: Behaviour of the functions $g(\alpha, \theta)$ (left) and $h(\alpha, \theta)$ (right), having a decisive role in determining indices of dispersion, for different values of parameter θ.

function $g(\alpha, \theta)$ becomes equal to zero, hence δ converges to the Poisson limit, where we have $\delta = 1$. Figure 7.2 exemplifies the ability of the above distribution to model overdispersed data and even more, to provide acceptable fits for empirical samples having $d \approx 1$, when commonly a 1-displaced Poisson model would be applicable. The performance of the CP distribution (blue dashed barplot) is compared to the Poisson distribution (red solid line) for two different situations, namely for $d \approx 1$ (left panel) and $d > 1$ (right panel). The original data represent Slovenian texts number 110 and 47, respectively, with observed absolute frequencies given in Table A.2, Appendix B. The corresponding parameter estimates are obtained by the maximum likelihood method, discussed in Section 7.4.2. In both cases the CP model gives a better fit.

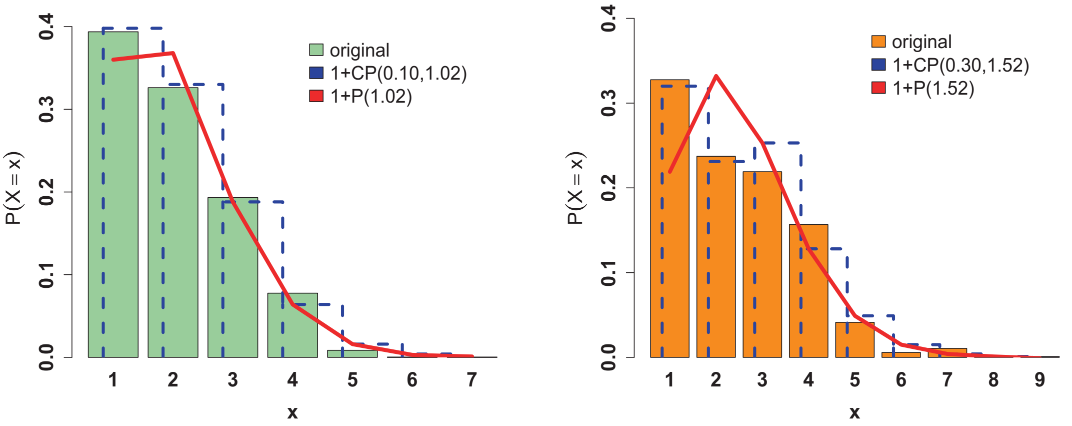


Figure 7.2: Examples of frequency distributions for equidispersed (left: $d = 1.01$) and overdispersed (right: $d = 1.28$) Slovenian text data, fitted by 1-displaced CP model, 1+CP(α, θ), and 1-displaced Poisson model, 1+P(θ).

7.3 Size-Biased Cohen-Poisson Distribution

Applying transformation (2.41) considered in Section 2.3.2 to the CP random variable X we obtain its size-biased version X^* with pmf defined by

$$\pi^*_{x|\alpha,\theta} = P(X^* = x) = \begin{cases} \dfrac{(1-\alpha)}{(1-\alpha e^{-\theta})}\, e^{-\theta}, & x = 1\,, \\ \dfrac{1}{(1-\alpha e^{-\theta})}\dfrac{\theta^{x-1}e^{-\theta}}{(x-1)!}, & x = 2, 3, \dots\,, \end{cases} \tag{7.7}$$

where $\theta > 0$ and $0 \le \alpha \le 1$. However, note that taking $\alpha^* = 1/(1-\alpha e^{-\theta})$ implies $1-\alpha^*+\alpha^* e^{-\theta} = (1-\alpha)e^{-\theta}/(1-\alpha e^{-\theta})$, and the above pmf corresponds to that of the 1-displaced SP distribution (4.2) with parameters α^* and θ, as defined in Chapter 4.

To derive the pgf of X^* we recall relation (2.42), Section 2.3.2 and substitute $G_X(t) = e^{\theta(t-1)} + \alpha\theta e^{-\theta}(1-t)$ and $\mathrm{E}(X) = \theta(1-\alpha e^{-\theta})$ from Table 7.2. Consequently, it follows that

$$G_{X^*}(t) = \frac{te^{-\theta}(e^{t\theta}-\alpha)}{1-\alpha e^{-\theta}}\,, \tag{7.8}$$

from which the mean and the variance are obtained as

$$\mu^* = \mathrm{E}(X^*) = 1 + \frac{\theta}{1-\alpha e^{-\theta}} \quad \text{and} \quad \mathrm{var}(X^*) = \frac{\theta}{1-\alpha e^{-\theta}} - \frac{\alpha e^{-\theta}\theta^2}{(1-\alpha e^{-\theta})^2}\,. \tag{7.9}$$

Subsequently, the factorial moments of the distribution (7.7) arise from its pgf as

$$\mu^*_{(1)} = 1 + \frac{\theta}{1-\alpha e^{-\theta}} \quad \text{and} \quad \mu^*_{(k)} = \frac{\theta^{k-1}(k+\theta)}{1-\alpha e^{-\theta}}\,, \ \text{ for } k \ge 2\,, \tag{7.10}$$

whereas raw and central moments can be determined from these applying the relations (B.9) and (B.13), Appendix B. Table 7.2 compares the original CP distribution to its 1-displaced and size-biased versions, by demonstrating changes in pmf, pgf and the first two moments of the given distributions.

Calculation of the size-biased CP probabilities is facilitated by the definition (7.7) for $\pi^*_{1|\alpha,\theta}$ and $\pi^*_{2|\alpha,\theta}$ and the following recurrence relation for the remaining ones

$$\pi^*_{x|\alpha,\theta} = \frac{\theta}{x-1}\pi^*_{x-1|\alpha,\theta}\,, \ \text{ for } x \ge 3\,. \tag{7.11}$$

Considering the index of dispersion given by

$$\delta = \frac{\mathrm{var}(X^*)}{\mathrm{E}(X^*)-1} = 1 - \frac{\alpha e^{-\theta}\theta}{1-\alpha e^{-\theta}} = 1 - \frac{g(\alpha,\theta)}{h(\alpha,\theta)}\,, \tag{7.12}$$

with $h(\alpha,\theta) = 2-\alpha e^{-\theta}$ and $g(\alpha,\theta)$ defined in (7.6), it can be shown that this distribution has the feature to model underdispersion with respect to Poisson variation. From the right panel of Figure 7.1 it is apparent that the function $h(\alpha,\theta)$ takes the

Table 7.2: Cohen-Poisson distribution: original, 1-displaced and size-biased forms

	Random Variable		
	X	X^d	X^*
Notation	CP(α, θ)	1+CP(α, θ)	1+SP(α^*, θ)
Range	$\mathbb{N}_0$	$\mathbb{N}$	$\mathbb{N}$
pmf	$\pi_{x\vert\alpha,\theta}$	$\pi_{x-1\vert\alpha,\theta}$	$\dfrac{x\pi_{x\vert\alpha,\theta}}{\theta(1-\alpha e^{-\theta})}$
pgf	$e^{\theta(t-1)} + \alpha\theta e^{-\theta}(1-t)$	$tG_X(t)$	$\dfrac{te^{-\theta}(e^{t\theta}-\alpha)}{1-\alpha e^{-\theta}}$
E$(\cdot)$	$\theta(1-\alpha e^{-\theta})$	1+E(X)	$1 + \dfrac{\theta}{1-\alpha e^{-\theta}}$
var$(\cdot)$	$\theta + \alpha\theta e^{-\theta}(2\theta - 1) - \alpha^2\theta^2 e^{-2\theta}$	var(X)	$\dfrac{\theta - \alpha e^{-\theta}(1+\theta)}{(1-\alpha e^{-\theta})^2}$

$\alpha^* = 1/(1-\alpha e^{-\theta})$

maximal value of 2 for $\alpha = 0$ or if θ is sufficiently large. Decreasing the value of θ, given a fixed value of α, will lead to a reduction in $h(\alpha, \theta)$. However, the function $h(\alpha, \theta)$ will never fall below unity, its lowest value is obtained when simultaneously $\alpha = 1$ holds and θ reaches its minimum. Consequently, since $h(\alpha, \theta)$ is a positive function for all parameter values allowed, the index of dispersion has its maximum at unity when $g(\alpha, \theta)$ becomes zero. Hence, the Poisson limit is attained for any α, whenever θ is large enough. The same occurs for $\alpha = 0$, irrespective of the value of θ. Table 7.3 illustrates the way in which underdispersion changes with alteration in α and θ. Obviously, as α increases, the value of δ decreases for any fixed value of θ. Figure 7.3 graphically illustrates the application of the size-biased CP model to two

Table 7.3: Size-biased Cohen-Poisson as adequate model choice for underdispersed data

		θ								
α		0.1	0.5	1	1.5	2	2.5	3	6	9
0	μ^*	1.10	1.50	2.00	2.50	3.00	3.50	4.00	7.00	10.00
	δ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	μ^*	1.12	1.57	2.08	2.57	3.06	3.54	4.03	7.00	10.00
	δ	0.98	0.93	0.92	0.93	0.94	0.96	0.97	1.00	1.00
0.4	μ^*	1.16	1.66	2.17	2.65	3.11	3.58	4.06	7.01	10.00
	δ	0.94	0.84	0.83	0.85	0.89	0.92	0.94	0.99	1.00
0.6	μ^*	1.22	1.79	2.28	2.73	3.18	3.63	4.09	7.01	10.00
	δ	0.88	0.71	0.72	0.77	0.82	0.87	0.91	0.99	1.00
0.8	μ^*	1.36	1.97	2.42	2.83	3.24	3.68	4.12	7.01	10.00
	δ	0.74	0.53	0.58	0.67	0.76	0.82	0.88	0.99	1.00
1	μ^*	2.05	2.27	2.58	2.93	3.31	3.72	4.16	7.01	10.00
	δ	0.05	0.23	0.42	0.57	0.69	0.78	0.84	0.99	1.00

Slovenian text samples, namely text no. 2 and text no. 89, with frequency distributions given in Table A.2, Appendix B. To calculate expected frequencies for the size-biased CP model, displayed in charts as blue dashed barplots, we use maximum likelihood parameter estimates and apply recurrence relation (7.11). Moreover, the comparison with the 1-displaced Poisson distribution as its lower limit, pictured with the red solid line, is also shown. Obviously, the size-biased CP distribution provides an adequate fit for empirical data with $d < 1$. For $d \approx 1$, both distributions fit well.

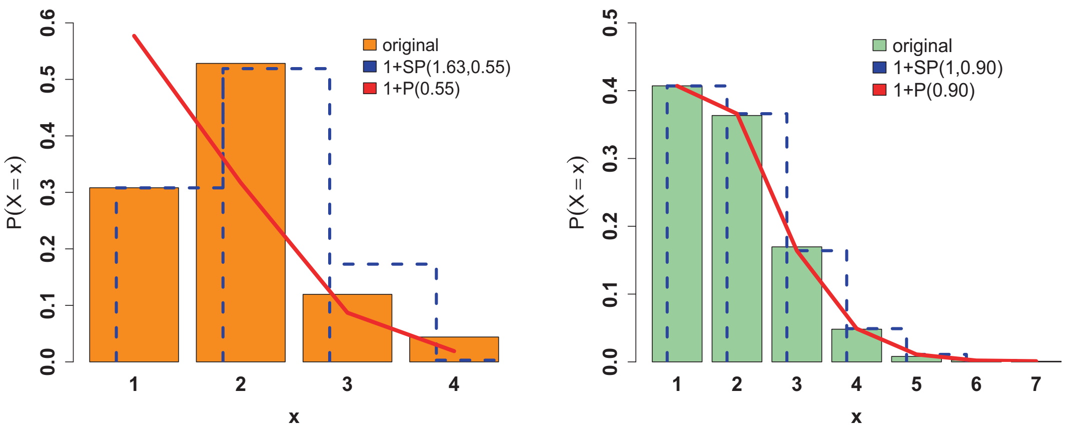


Figure 7.3: Examples of frequency distributions for underdispersed (left: $d = 0.66$) and equidispersed (right: $d = 0.99$) Slovenian text data, fitted by size-biased CP model, 1+SP(α^*, θ), where $\alpha^* = 1/(1 - \alpha e^{-\theta})$, and 1-displaced Poisson model, 1+P(θ).

7.4 Parameter Estimation

In this section we derive parameter estimators for the 1-displaced CP model (7.1) and the size-biased CP model (7.7) based on the three methods introduced in Section 2.4. However, since the size-biased CP(α, θ) distribution corresponds to the 1-displaced SP(α^*, θ) distribution, with $\alpha^* = 1/(1 - \alpha e^{-\theta})$, estimators of the parameter θ for model (7.7) are identical to that of model (4.2) given in Section 4.4, for all three techniques applied.

7.4.1 Estimation by Method of Moments

By equating the sample mean $\bar{x}$ and the second factorial sample moment $m_{(2)}$ to the theoretical mean $\mu^d = 1 + \theta(1 - \alpha e^{-\theta})$ and the second factorial theoretical moment $\mu^d_{(2)} = \theta^2 + 2\theta(1 - \alpha e^{-\theta})$, respectively, we get the moment estimators of the parameters α and θ, for the *1-displaced CP model*, defined by (7.1), as follows

$$\hat{\theta}_{\mathrm{MM}} = \sqrt{m_{(2)} - 2(\bar{x} - 1)} \quad \text{and} \quad \hat{\alpha}_{\mathrm{MM}} = \frac{\hat{\theta}_{\mathrm{MM}} - \bar{x} + 1}{\hat{\theta}_{\mathrm{MM}}\, e^{-\hat{\theta}_{\mathrm{MM}}}}\,. \tag{7.13}$$

To obtain the corresponding moment estimators for the *size-biased CP model*, given in (7.7), we proceed analogously as above. Solving simultaneously the following system of equations

$$\bar{x} = 1 + \frac{\theta}{1-\alpha e^{-\theta}} \quad \text{and} \quad \mu_{(2)} = \frac{\theta(2+\theta)}{1-\alpha e^{-\theta}}\,,$$

we get the following moment estimators

$$\hat{\theta}_{\text{MM}} = \frac{m_{(2)}}{\bar{x}-1} - 2 \quad \text{and} \quad \hat{\alpha}_{\text{MM}} = \frac{\hat{\theta}_{\text{MM}} - \bar{x} + 1}{(1-\bar{x})e^{-\hat{\theta}_{\text{MM}}}}\,. \tag{7.14}$$

It should be noted here that the moment estimator $\hat{\theta}_{\text{MM}}$ of θ for the size-biased CP model coincides with that of the 1-displaced SP model (4.2), given in equation (4.13).

7.4.2 Estimation by Maximum Likelihood

The likelihood function of a random sample of size n from the *1-displaced CP model* with the pmf (7.1) is given by

$$L(\alpha,\theta|f_1,\ldots,f_k) = \left((1+\alpha\theta)e^{-\theta}\right)^{f_1}\left((1-\alpha)\theta e^{-\theta}\right)^{f_2}\prod_{i=3}^{k}\left(\frac{\theta^{i-1}e^{-\theta}}{(i-1)!}\right)^{f_i}. \tag{7.15}$$

Maximum likelihood estimating equations obtained by equating to zero the partial derivatives of $l(\alpha,\theta|f_1,\ldots,f_k) = \log L(\alpha,\theta|f_1,\ldots,f_k)$ with respect to α and θ are

$$\frac{\partial l(\alpha,\theta|f_1,\ldots,f_k)}{\partial\alpha} = \frac{f_1\theta}{1+\alpha\theta} - \frac{f_2}{1-\alpha} = 0\,, \tag{7.16}$$

$$\frac{\partial l(\alpha,\theta|f_1,\ldots,f_k)}{\partial\theta} = \frac{f_1\alpha}{1+\alpha\theta} + \frac{f_2}{\theta} + \sum_{i=3}^{k}\frac{f_i(i-1)}{\theta} - \sum_{i=1}^{k} f_i = 0\,. \tag{7.17}$$

This equations can be simplified as follows

$$\alpha = \frac{f_1\theta - f_2}{\theta(f_1+f_2)} \quad \text{and} \quad \frac{f_1\alpha}{1+\alpha\theta} + \frac{n(\bar{x}-\theta-1)}{\theta} = 0\,. \tag{7.18}$$

We substitute the expression above for α in the second equation of (7.18) to obtain the estimator $\hat{\theta}_{\text{ML}}$ as a solution of the following equation

$$\theta^2 - \left(\bar{x} - 2 + \frac{f_1}{n}\right)\theta - \left(\bar{x} - 1 - \frac{f_2}{n}\right) = 0\,, \tag{7.19}$$

which is quadratic in θ and has two roots in cases where estimates exist. The positive root is the required estimator and is given by the quadratic formula as

$$\hat{\theta}_{\text{ML}} = \frac{1}{2}\left(\bar{x} - 2 + \frac{f_1}{n}\right) + \frac{1}{2}\sqrt{\left(\bar{x} - 2 + \frac{f_1}{n}\right)^2 + 4\left(\bar{x} - 1 - \frac{f_2}{n}\right)}\,. \tag{7.20}$$

Then, the estimator $\hat{\alpha}_{\text{ML}}$ follows from the first equation of (7.18) by inserting $\hat{\theta}_{\text{ML}}$.

The likelihood function of the *size-biased CP model* with pmf (7.7) is

$$L(\alpha,\theta|f_1,\ldots,f_k) = \left(\frac{(1-\alpha)e^{-\theta}}{1-\alpha e^{-\theta}}\right)^{f_1} \prod_{i=2}^{k} \left(\frac{\theta^{i-1}e^{-\theta}}{(1-\alpha e^{-\theta})(i-1)!}\right)^{f_i}. \tag{7.21}$$

We take logarithms of (7.21), differentiate with respect to α and θ, equate to zero and subsequently simplify to get the following ML estimating equations

$$\frac{\partial l(\alpha,\theta|f_1,\ldots,f_k)}{\partial\alpha} = \frac{-f_1}{1-\alpha} + \frac{ne^{-\theta}}{1-\alpha e^{-\theta}} = 0\,, \tag{7.22}$$

$$\frac{\partial l(\alpha,\theta|f_1,\ldots,f_k)}{\partial\theta} = \frac{n(\bar{x}-1)}{\theta} - n - \frac{n\alpha e^{-\theta}}{1-\alpha e^{-\theta}} = 0\,. \tag{7.23}$$

Further substituting the expression for α, derived directly from (7.23), in equation (7.22) we obtain the estimator $\hat{\theta}_{\text{ML}}$ of θ as a solution of the following equation

$$\frac{\theta(n-f_1)}{n(\bar{x}-1)} + e^{-\theta} - 1 = 0\,. \tag{7.24}$$

The solution of this transcendental equation can be found by the same procedure as outlined for the solution of equation (4.19) in Chapter 4, since the two equations coincide. Hence, the estimator $\hat{\alpha}_{\text{ML}}$ is obtained by substituting $\hat{\theta}_{\text{ML}}$ in the expression for α. Note that the resulting $\hat{\alpha}_{\text{ML}}$ is analogous to that of the method of moments given in (7.14), the only difference being that $\hat{\theta}_{\text{MM}}$ is replaced here by $\hat{\theta}_{\text{ML}}$.

7.4.3 Estimation Based on Mean and First Frequency Class

The estimators for the parameters of the *1-displaced CP model* (7.1) based on mean and first frequency class, denoted here by $\hat{\alpha}_{\text{FF}}$ and $\hat{\theta}_{\text{FF}}$, result from the system of simultaneous equations given by

$$\bar{x} = 1 + \theta(1-\alpha e^{-\theta}) \quad \text{and} \quad f_1/n = (1+\alpha\theta)e^{-\theta}\,. \tag{7.25}$$

Solving the first equation of (7.25) we get the same expression for the estimator $\hat{\alpha}_{\text{FF}}$ of α as that of $\hat{\alpha}_{\text{MM}}$ given in (7.13), where instead of $\hat{\theta}_{\text{MM}}$ we have $\hat{\theta}_{\text{FF}}$. To derive the estimator $\hat{\theta}_{\text{FF}}$ of θ we substitute the expression for α given by the first equation of (7.25) in the second one. Consequently, after a few algebraic modifications $\hat{\theta}_{\text{FF}}$ follows as a solution of the following equation

$$\theta + e^{-\theta} + 1 - \bar{x} - f_1/n = 0\,. \tag{7.26}$$

To obtain the corresponding estimators of the *size-biased CP model* (7.7) we search for the simultaneous solution of the following estimating equations

$$\bar{x} = 1 + \frac{\theta}{1-\alpha e^{-\theta}} \quad \text{and} \quad \frac{f_1}{n} = \frac{(1-\alpha)e^{-\theta}}{1-\alpha e^{-\theta}}\,. \tag{7.27}$$

Applying the same algebraic approach leading to equation (7.26), it can be easily shown that the parameter estimators $\hat{\theta}_{\text{FF}}$ and $\hat{\alpha}_{\text{FF}}$ are identical to the corresponding ML estimators $\hat{\theta}_{\text{ML}}$ and $\hat{\alpha}_{\text{ML}}$, given in Section 7.4.2.

7.5 A Simulation Study

Wanting to further investigate the behavior of the proposed estimation procedures we carry out a simulation experiment based on the results obtained for representative Slovenian journalistic texts, private letters/prose, and poems. These representatives are created as aggregation of texts under study belonging to the same genre. Being aware of the fact that the journalistic data sets are overdispersed, the majority of the private letters and prose texts are equidispersed, whereas poems are underdispersed, we take $(\alpha, \theta) = (0.29, 1.38)$ and $(\alpha, \theta) = (0.11, 0.90)$ to generate the overdispersed $(\delta > 1)$ and equidispersed $(\delta = 1)$ 1-displaced CP samples, respectively. These choices coincide with the corresponding ML estimates of the given text type aggregations. In order to analyze the underdispersed $(\delta < 1)$ data case we create the size-biased CP sample with ML estimates of the representative poem, given by $(\alpha, \theta) = (0.20, 0.66)$, taken as model parameters. However, neither for private letters nor for prose texts aggregation, the ML results obtained were plausible when fitting this model. Since for $\alpha = 0$ the CP models simplify to Poisson, we set $(\alpha, \theta) = (0, 0.90)$ to study performance of the estimation approaches near the boundary case. To generate 1-displaced and size-biased CP random variables we apply the inversion method, introduced in Section 2.5, where the corresponding probabilities are calculated using the recurrence formula (7.5) and (7.11), respectively. For all situations mentioned above we create $M = 500$ Monte Carlo samples of size $n = 500$ and $n = 1000$. Tables 7.4 and 7.5 present the estimation results obtained. The mean values of $M = 500$ parameter estimates, denoted here by $\bar{\alpha}$ and $\bar{\theta}$, based on method of moments (MM), maximum likelihood (ML) and that of mean and first frequency class (FF) are calculated. Also, the estimated standard errors of the mean values, denoted by $\mathrm{se}_{\bar{\alpha}}$ and $\mathrm{se}_{\bar{\theta}}$, are obtained for each estimation technique. These are calculated as the standard deviation of the $M = 500$ parameter estimates.

As it is evident from Table 7.4, relating to the 1-displaced CP model, the best results are those of the ML method, independently of the analyzed dispersion situation. Also, the standard errors of the estimated parameters are the smallest here and decrease with increasing sample size. However, for $\delta = 1$ we could not obtain all estimation results. For the sample of size $n = 500$ we got 444 valid MM esti-

Table 7.4: Estimation results for over- and equidispersed data situations

	$(\alpha,\theta) = (0.29, 1.38)$					
$\delta > 1$	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\alpha}_{\mathrm{FF}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.283 (0.099)	1.378 (0.061)	0.286 (0.051)	1.379 (0.054)	0.285 (0.057)	1.379 (0.055)
$n = 1000$	0.282 (0.071)	1.376 (0.043)	0.285 (0.036)	1.377 (0.039)	0.284 (0.040)	1.377 (0.039)
	$(\alpha,\theta) = (0.11, 0.90)$					
$\delta = 1$	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	$\bar{\alpha}_{\mathrm{FF}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{FF}}})$	$\bar{\theta}_{\mathrm{FF}}(\mathrm{se}_{\bar{\theta}_{\mathrm{FF}}})$
$n = 500$	0.124 (0.075)	0.907 (0.051)	0.117 (0.053)	0.905 (0.047)	0.113 (0.057)	0.904 (0.048)
$n = 1000$	0.112 (0.058)	0.902 (0.037)	0.110 (0.040)	0.901 (0.034)	0.101 (0.043)	0.900 (0.034)

mates, whereas there were 442 and 477 available results in the case of ML and FF estimation, respectively. Increasing the sample size by factor 2, we still had 23 not obtainable MM estimates, 24 of the ML method and 5 of the FF method. Figure 7.4 illustrates 95% and 99% confidence ellipses of the corresponding parameter pairs for the sample of size $n = 1000$ and the $\delta > 1$ data situation. In all three cases, the coverage of the true values is apparent.

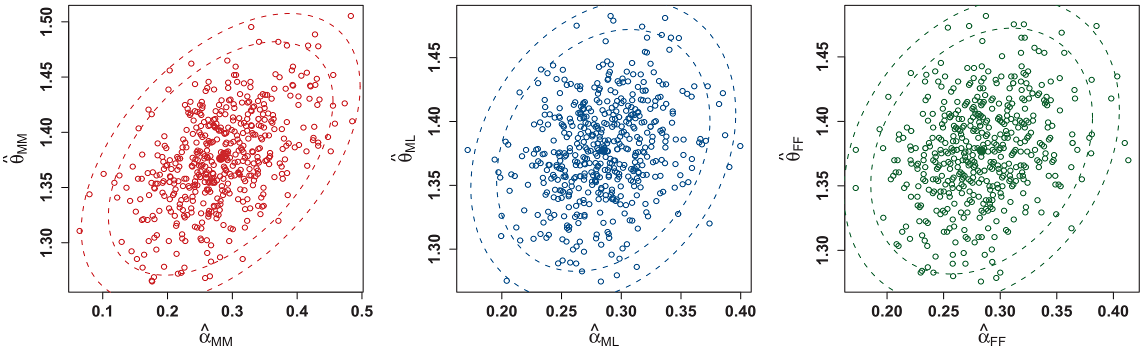


Figure 7.4: 95% and 99% confidence ellipses for overdispersed data sample of size 1000

Estimation results for the size-biased CP model are displayed in Table 7.5 below. As FF estimates of the size-biased model coincide with those of ML, they are documented only once. The number of samples resulting in valid estimates are given in columns named M_{valid}. Note that when $\delta = 1$, approximately 50% of all M=500 possible results are reported only. Once again, slightly better results are obtained for ML estimates, regardless of the observed dispersion data situation. The standard errors are again smaller in this case. However, as apparent from Figure 7.5 parameter estimates of α and θ are highly correlated for both estimation methods. Thereby, the estimated correlation coefficient of the M=500 replications is computed

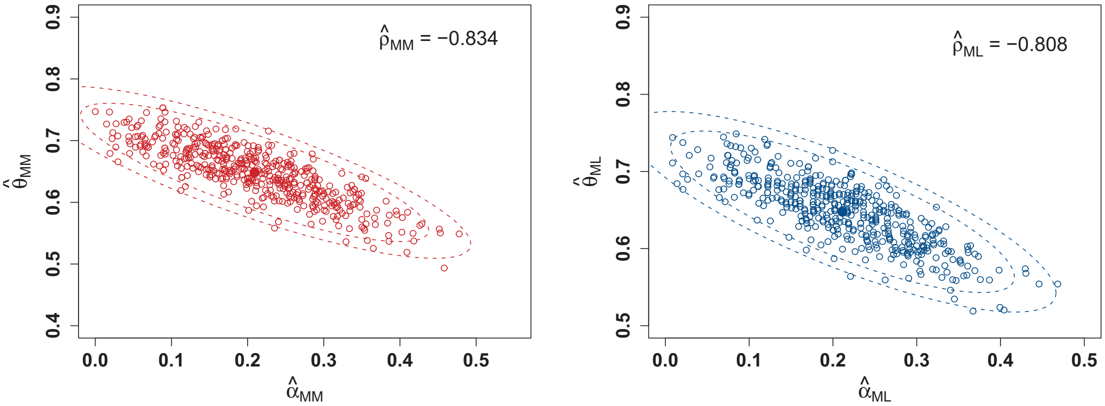


Figure 7.5: 95% and 99% confidence ellipses for underdispersed data sample of size 1000

as $\widehat{\rho} = cov(\hat{\alpha}, \hat{\theta})/\mathrm{se}_{\bar{\alpha}}\mathrm{se}_{\bar{\theta}}$. The correlations of the $n = 500$ generated data points with M=500 replications are $\widehat{\rho}_{\mathrm{MM}} = -0.827$ and $\widehat{\rho}_{\mathrm{ML}} = -0.811$, whereas for $n = 1000$ we have $\widehat{\rho}_{\mathrm{MM}} = -0.834$ and $\widehat{\rho}_{\mathrm{ML}} = -0.808$.

Table 7.5: Estimation results for under- and equidispersed data situations

	$(\alpha, \theta) = (0.20, 0.66)$					
$\delta < 1$	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	M_{valid}	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	M_{valid}
$n = 500$	0.232 (0.114)	0.639 (0.062)	440	0.230 (0.105)	0.640 (0.058)	432
$n = 1000$	0.209 (0.093)	0.648 (0.046)	478	0.211 (0.084)	0.648 (0.043)	475
	$(\alpha, \theta) = (0, 0.9)$					
$\delta = 1$	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	M_{valid}	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$	M_{valid}
$n = 500$	0.127 (0.085)	0.850 (0.054)	267	0.122 (0.076)	0.851 (0.052)	220
$n = 1000$	0.094 (0.069)	0.863 (0.041)	264	0.085 (0.060)	0.865 (0.037)	235

Chapter 8

Application to Data

8.1 Introduction

In this chapter the two-parametric generalizations of the Poisson model considered in the previous chapters are applied to text data of different genres from two particular Slavic languages. The main interest of this study is to find a unique model for word length frequency distributions of texts selected that covers the whole d range. Then, the totality of all texts of a given natural language would be an extreme realization of this procedure. We look at the goodness of fit properties of the analyzed models and check the suitability of the estimated model parameters to discriminate the text sorts given. This is done with the aim of getting the information whether different texts from a single language but also across languages can be compared and distinguished on the basis of specific model parameters.

8.2 Data Base of the Study

The 120 Slovenian, as well as Russian texts, serving as a basis for this study represent four different text types, namely *journalistic*, *poems*, *private letters* and *prose*, thirty texts of each text type being analyzed.[1] These texts have been systematically chosen based on findings from recent word length studies (cf. Antić et al., 2006a; Grzybek et al., 2005b) in order to generate a *balanced experimental design* with equal number of observations for each text group. As to historical and methodological background of these studies see the contribution of Grzybek (2006). The specific selection of the texts sorts above has been deliberately made to cover the broad textual spectrum, or at least its extreme realizations and relies on the scheme of Figure 1.1, Section 1.3. In accordance with the considerations of Section 1.3, homogeneous texts or parts of texts are taken as analytical units. The detailed reference for the texts under study are given in Appendix A, Tables A.1 for the Slovenian and A.7 for the Russian

[1] The text basis of this study is part of the text data base developed in the Graz research project "Word Length Frequencies in Slavic Texts", mentioned already.

language. Frequency distributions with the first two empirical moments and dispersion index are summarized in Tables A.2 and A.8, respectively. Table 8.1 illustrates the characteristical statistical measures for our text sample: mean word length ($\bar{x}$), sample variance (s^2), text length (TL) and index of dispersion (d). Although not very different with respect to mean word length and sample variance, Slovenian prose texts seem to be more dispersed than private letters regarding text length as Figure 8.1, left panel, clearly shows. Evidently, the shortest are the poems, followed by journalistic texts which are, however, characterized by the longest words of up to nine syllables. As opposed to this, it turned out that among all Russian texts observed, the private letters are the shortest ones, though having the widest range of 1.83 up to 2.52 syllables per word on average, as shown in Table 8.1. Again, the longest words are specific for journalistic texts, whereas prose texts are the most

Table 8.1: Statistical measures of Slovenian and Russian texts under study

			$\bar{x}$		s^2		TL		d	
	Text Type	N	min	max	min	max	min	max	min	max
Slovenian	Journalistic	30	2.05	2.46	1.22	1.96	328	1166	1.09	1.35
	Poems	30	1.48	1.90	0.37	0.84	58	626	0.60	1.14
	Private letters	30	1.72	1.98	0.78	0.98	401	1979	0.97	1.15
	Prose	30	1.73	1.98	0.70	1.04	288	4401	0.95	1.16
Russian	Journalistic	30	2.40	2.83	1.46	2.17	320	901	1.03	1.34
	Poems	30	1.76	2.40	0.73	1.60	77	1014	0.76	1.41
	Private letters	30	1.83	2.52	0.90	2.11	48	488	0.79	1.53
	Prose	30	2.02	2.52	1.15	1.83	236	3154	0.93	1.24

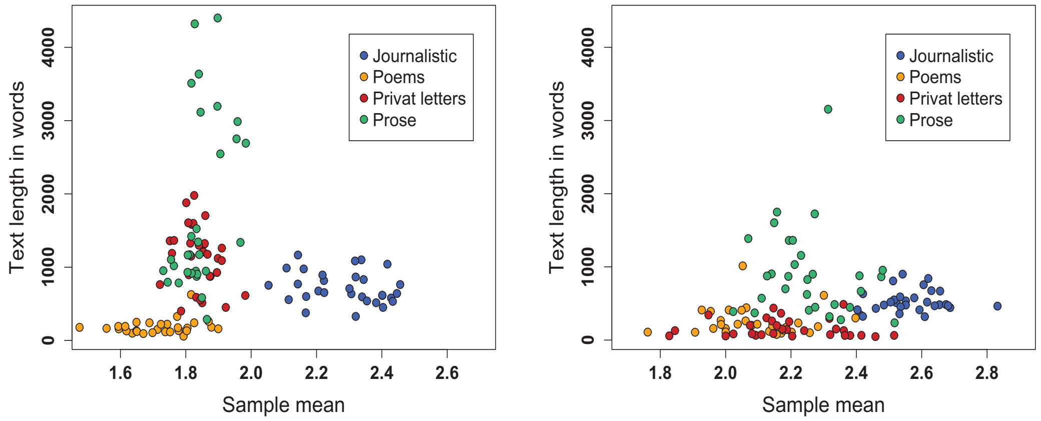


Figure 8.1: Differences in text length measured as the number of words and word length measured as the number of syllables regarding text type for Slovenian (left) and Russian (right) texts under study.

dispersed with respect to text length, as demonstrated in Figure 8.1, right panel. Figure 8.2 visualizes the differences in the mean word length of the Slovenian texts compared to that of the Russian texts for four text types given. Evidentally, the average word lengths in Russian texts are significantly longer than in Slovenian texts (Wilcoxon rank-sum test, p<0.01). The detailed theoretical explanation regarding the question of the linguistic differences between the Slovenian and Russian language can not be given here. However, analyzing the Slavic parallel corpus, Kelih (2009b) has reached the same conclusion. He mentioned, among other reasons, that the Russian compared to Slovenian has more complex syllable structure, and on the morphological level tends to form longer word forms, in particular compound words. For details see also Kelih (2011).

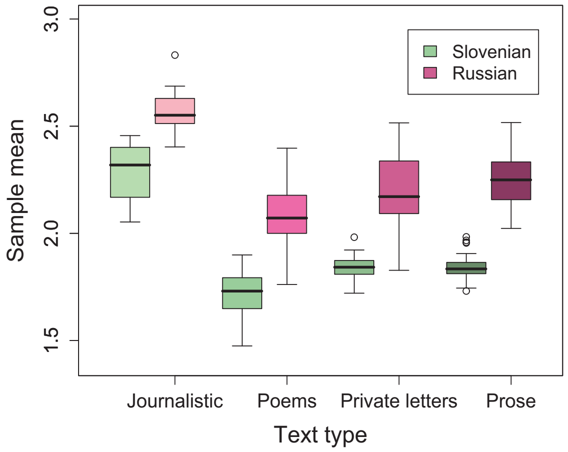


Figure 8.2: Differences in word length between languages for different text types.

The left panel of Figure 8.3 shows results of plotting sample mean versus sample variance for all 120 Slovenian texts. The solid black line is the Poisson reference

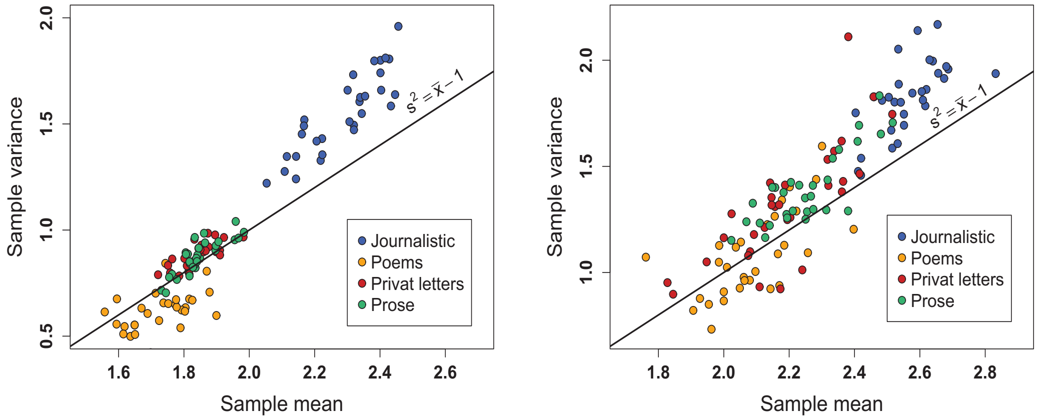


Figure 8.3: Sample mean compared to sample variance for Slovenian (left) and Russian (right) texts selected with respect to different text types.

line where the sample mean diminished by one is near to the sample variance, i.e. the equidispersion holds. Obviously, all journalistic texts lie above the Poisson line, private letters and prose texts are around this line, while most of the poems lie below the Poisson line. Being aware of this fact and based on findings in Section 2.1, we calculated the index of dispersion for each of 120 Slovenian texts selected. It turned out that $d > 1$ for all 30 journalistic texts. 27 out of 30 Slovenian poems have $d < 1$, whereas for private letters and prose texts we obtain $0.98 < d < 1.05$ in 43.33% and 63.33% cases, respectively. Concerning Russian texts the right panel of Figure 8.3 clearly displays that all journalistic texts are overdispersed, just like the most of the private letters (76.67%) and prose texts (80%), while for poems under-, equi- and overdispersion situations may come into question. Figure 8.4 graphically displays our obtained results regarding index of dispersion for both Slavic languages.

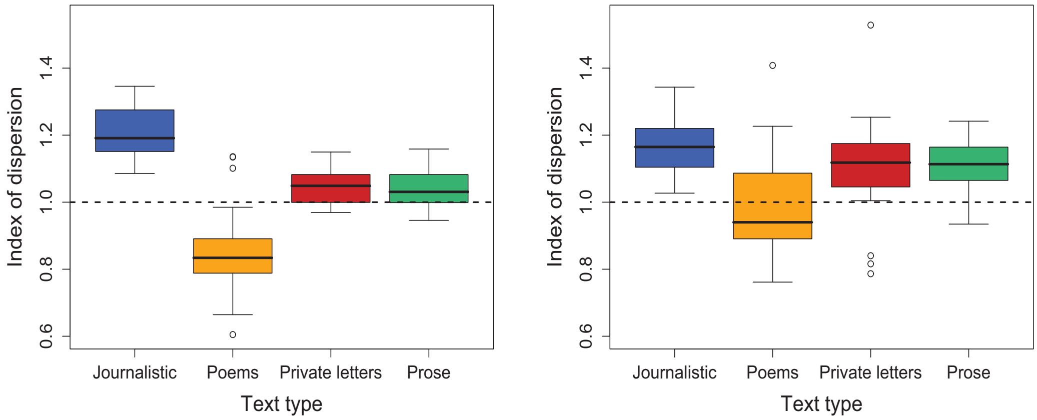


Figure 8.4: Differences in index of dispersion regarding text type for Slovenian (left) and Russian (right) texts selected

8.3 Model Selection

As an essential result of the considerations in the previous chapters it turned out that due to the specific syllable and word structure in the analyzed Slavic languages several two-parametric probability models can be taken into account for the word length frequency distributions of the texts under study. Nevertheless, the final decision depends on whether the chosen model has ability to cover the whole d range, and thus allows for modelling under-, equi- and overdispersed count data. Table 8.2 summarizes our findings for 1-displaced and size-biased versions of the discussed models. Because $\delta \leq 1$ for the 1-displaced Dacey-Poisson and Kemp-Kemp-Poisson distributions, they are likely to be adequate only for empirical samples with under- or equidispersion.[2] Quite contrary to this, having $\delta > 1$ both versions of the nega-

[2] The size-biased versions of Dacey-Poisson and Kemp-Kemp-Poisson distributions were not the scope of this study, hence are marked as grey areas in Table 8.2.

tive binomial models offer possible solutions in the case of overdispersion. The same holds for the 1-displaced Cohen-Poisson distributions. Interestingly enough, its size-biased competitor is suitable exclusively for count data being underdispersed. This is the reason why we obtained unreliable estimates $\hat{\alpha}$ for poems in the case of the 1-displaced CP model, but get the plausible one if applying its size-biased version (cf. Table A.6 and A.12, Appendix A). The generalized Poisson distributions, approximating both binomial and negative-binomial distributions, allow for all types of dispersion. However, its 1-displaced version proved to be more advantageous than the size-biased one. Some convincing arguments are the following: (i) based on the parameter λ it is possible to clearly determine the type of mean-variance relationship, and thus to identify Poisson under-/overdispersion as well as the Poisson limit, (ii) the parameter estimates are stable, easy obtainable and implementable, (iii) simulation studies show good performance of the model for all dispersion situations. Moreover, standard errors of $\bar{\theta}$ proved to be smaller compared to those of the size-biased case (cf. Section 6.5). Although providing an unified approach for all data cases, detailed analysis showed that both Hyper-Poisson models are unsuitable for the Slovenian journalistic texts and poems. The explanation for this fact may lie in the structure of the journalistic texts displaying almost the same frequencies of two- and three-syllable words. However, a satisfactory fit requires instead a monotonic decreasing trend of these frequencies, as demonstrated in Antić et al. (2006b). As to Russian texts, it turned out to be rather problematic to fit some of the private letters and poems. Finally, the 1-displaced Singh-Poisson model providing simple alternative to the Poisson distribution is applicable in situations where the observed count data have $d \neq 1$, but has also ability for modelling equidispersion. In all cases we obtain reasonable and stable estimates of the model parameters. Impressions from real data analysis are also supported by the simulation study (cf. Section 4.5). The special benefit of this model is that the maximum likelihood estimation pro-

Table 8.2: Index of dispersion as crucial value for real data

	1-displaced			*size-biased*		
Distribution	$\delta < 1$	$\delta \approx 1$	$\delta > 1$	$\delta < 1$	$\delta \approx 1$	$\delta > 1$
Binomial	✓			✓		
Dacey-Poisson	✓	✓				
Generalized Poisson	✓	✓	✓	✓	✓	✓
Hyper-Poisson	✓	✓	✓	✓	✓	✓
Kemp-Kemp-Poisson	✓	✓				
Negativ-Binomial			✓			✓
Poisson		✓				
Singh-Poisson	✓	✓	✓		✓	
Cohen-Poisson		✓	✓	✓	✓	

cedure leads to the same estimates as the method based on the sample mean and the first frequency class. Since it is primarily applicable when higher proportion of ones is at hand, we analyze in the next step which percentage of the whole text corpus (i.e. all sampled texts) is represented by x-syllable words, for Slovenian and Russian language separately. The results of the same analysis but with respect to each of the four text types are represented in Table 8.3. Figure 8.5, left panel, clearly shows that for Slovenian texts the percentage of one-syllable words is the highest, both as compared to the whole text corpus, and to isolated samples of the four text groups mentioned above (see the upper part of Table 8.3). However, the difference between one- and two-syllable words is not so obvious when Russian texts are considered. Even more, for the journalistic texts we notice rather the decrease of one-syllable words compared to the other text types. Consequently, we restrict our

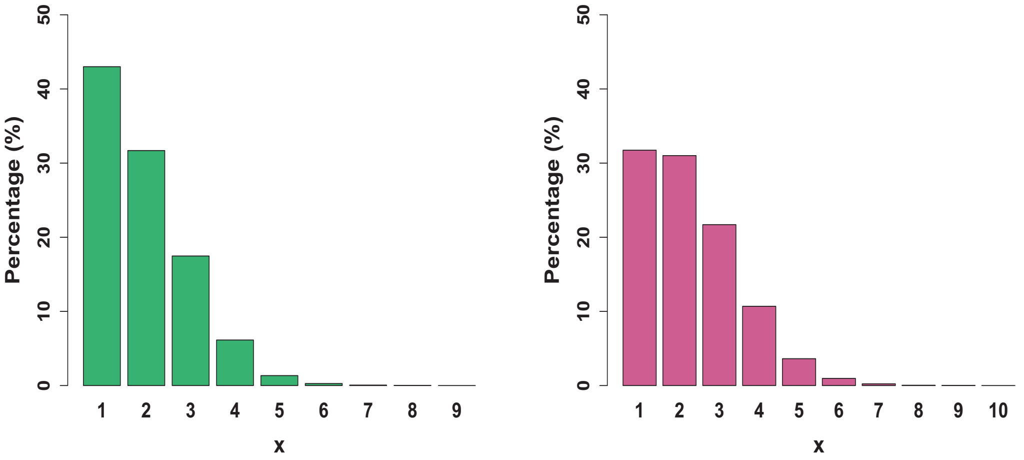


Figure 8.5: Word length frequencies (in %) in Slovenian (left) and Russian (right) corpus

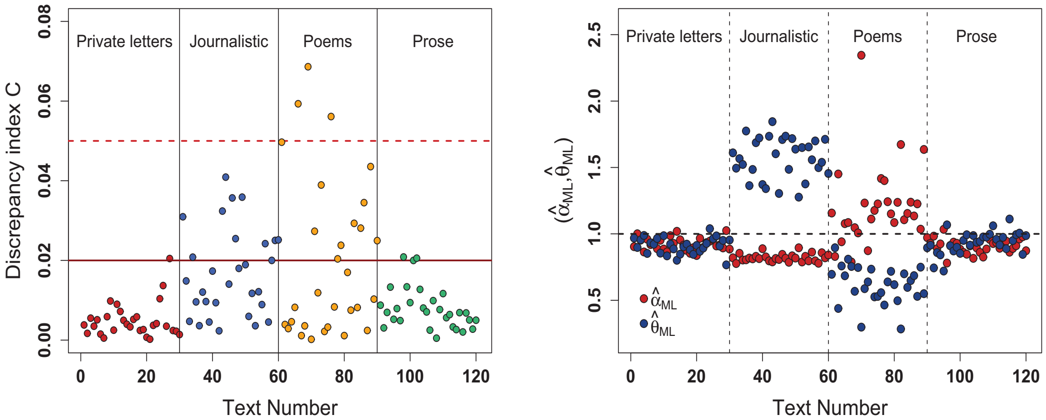
Table 8.3: Percentage of x-syllable words for Slovenian and Russian texts under study

		x-syllable words									
	Text Type	1	2	3	4	5	6	7	8	9	10
Slovenian	Journalistic	35.21	24.72	22.71	12.39	3.77	0.90	0.23	0.06	0.01	-
	Poems	46.69	36.75	14.62	1.87	0.08	0.00	0.00	0.00	0.00	-
	Private letters	44.87	33.39	15.66	5.10	0.84	0.13	0.01	0.00	0.00	-
	Prose	45.25	31.90	16.87	5.22	0.69	0.06	0.00	0.00	0.00	-
Russian	Journalistic	25.60	28.05	23.38	13.78	6.38	2.11	0.56	0.09	0.05	0.01
	Poems	34.82	34.30	20.21	8.48	2.02	0.16	0.00	0.00	0.00	0.00
	Private letters	34.32	30.84	20.93	10.44	2.62	0.65	0.16	0.05	0.00	0.00
	Prose	32.21	30.85	22.27	10.07	3.48	0.91	0.18	0.03	0.02	0.00

further analysis to *1-displaced generalized Poisson* and *Singh-Poisson distributions* only. Furthermore, it is of particular interest to find out whether the chosen model(s) allow for text discrimination. Most studies in this field, particulary the ones mentioned above, analyzed the relevance of word length in general, and for purpose of text classification. Based on the methods of multivariate analysis, it has been shown that word length, if properly defined as the number of syllables per word, is a characteristic of genre, rather than a variable describing author's individual style, thus playing a crucial role in the discrimination and classification of text types (cf. Kelih et al., 2005; Stadlober and Djuzelic, 2006). In these studies, discriminant analysis uses mainly characteristics derived from the empirical frequency distributions. Here, however, we carry out discriminant analysis based on the parameters of theoretical discrete probability models fitted to the observed data.

8.4 Interpretation of Parameters

The results of fitting the Singh-Poisson model to 120 Slovenian texts are displayed in Figure 8.6a. The solid red line in the graph is the reference bound $C = 0.02$, whereas the dashed line refers to $C = 0.05$. Obviously, the SP model provides a good fit for the majority of the texts. It seems not to be appropriate just for the three poems of Gregorčić, namely "Kesanje" (text no. 66), "Na sveti večer" (text no. 69) and "Pri zibelki" (text no. 76). A closer look at their structure shows that all of them are indeed short texts, having length of 181, 117 and 114 words, respectively. The natural question arising from this fact is whether and how goodness of fit may be influenced by text length, and in particular if a minimal length of texts is needed in order to achieve satisfactory results. Kelih (2012) investigated in more detail the question of a "*minimal*", "*maximal*" and "*optimal*" text length, with a specific



(a) Discrepancy index

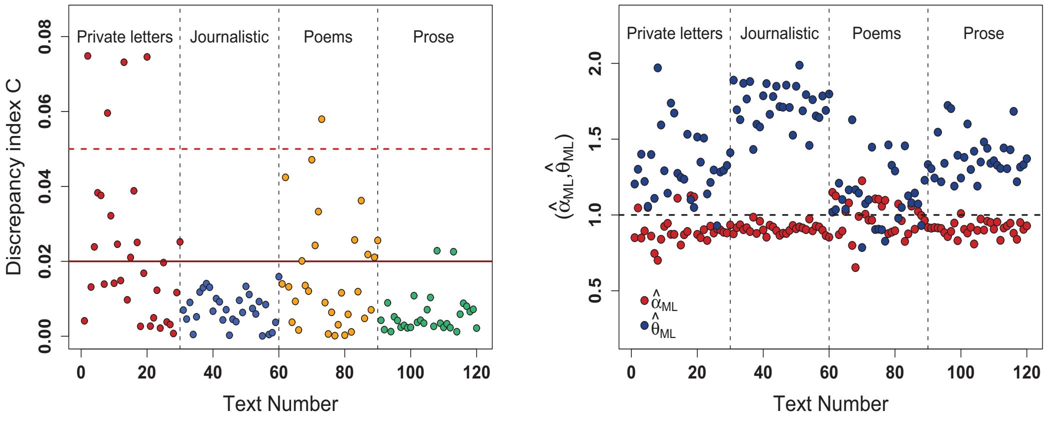
(b) $(\hat{\alpha}_{\text{ML}}, \hat{\theta}_{\text{ML}})$ parameter regions

Figure 8.6: Results of fitting Singh-Poisson model to 120 Slovenian texts

accent on its interaction with word length. The main result of his study says that these two basic text properties are directly related (the longer the text, the longer its words), and can be explained by an appropriate mathematical model, at least for the Russian and Bulgarian texts studied. This means that at the beginning of the text the author uses some "known" vocabulary. But, being forced to continually provide a flow of new information he starts to involve more infrequent and unusual words, generally longer than the frequent ones, hence text begins to grow systematically. Short texts, like those above, still possibly miss this kind of "balance" as to the use of differently long words. Besides, Best and Zinenko (1998) argued that only letters can be accepted as a representative of "natural" short texts, since they are written in one act of creating written works, and hence present an "optimal" text sort. A text containing *at least 250 words* corresponds thus to "natural" act of speaking (Best, personal communication); the three poems above not requiring this pre-condition.

For all 120 Slovenian texts maximum likelihood (ML) estimates of both parameters were computed and each pair of parameters $(\hat{\alpha}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}})$ was plotted versus the corresponding text, as shown in Figure 8.6b. The estimated parameters $\hat{\alpha}_{\mathrm{ML}}$ are represented by red circles, while estimated parameters $\hat{\theta}_{\mathrm{ML}}$ are signified by the blue ones. It is evident that each group of texts leads to a different pattern of parameters. In case of private letters both parameters are very close to each other, the same holds for prose texts although reversed in respect to the order. Contrary to this, in journalistic texts and poems parameters are quite distant from each other. The $\hat{\alpha}_{\mathrm{ML}}$ outlier in Figure 8.6b refers to Gregorčić's poem "Njega ni!" (text no. 70). This text has only 106 words, $\hat{\alpha}_{\mathrm{ML}} = 2.35$ and $\hat{\theta}_{\mathrm{ML}} = 0.3$ ($\alpha_{\max} = 3.88$). Surprisingly, the value of $C = 0.0002$ indicates here rather an extremely good fit.

The results of fitting the Singh-Poisson model to 120 Russian texts are displayed in Figure 8.7. Notice that for five texts the values of C are beyond the reference line. These texts include again four extremely short private letters of Achmatova



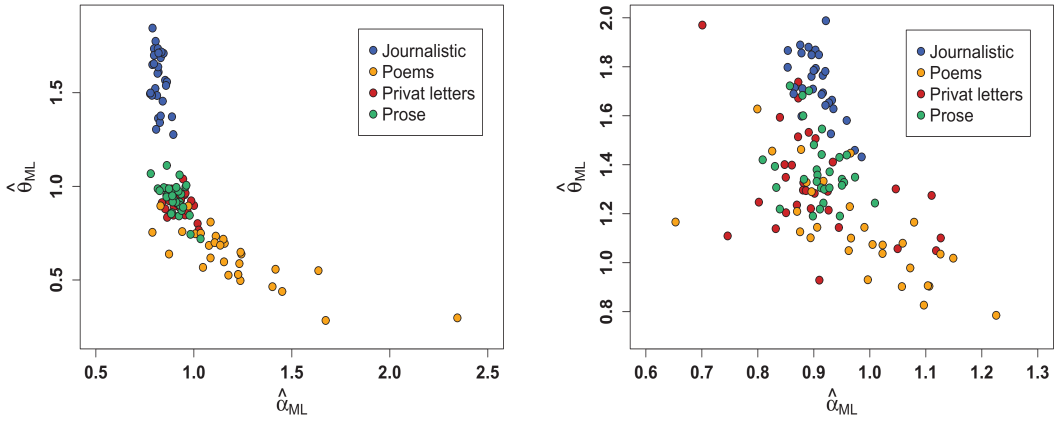
(a) Discrepancy index

(b) $(\hat{\alpha}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}})$ parameter regions

Figure 8.7: Results of fitting Singh-Poisson model to 120 Russian texts

written to Brodsky, Chardžiev and Maksimov (texts no. 2, 8, 13 and 20) consisting of 130, 63, 48 and 75 words, respectively, and poem of Nekrasov "Muza" (text no. 73) with 302 words. The $(\hat{\alpha}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}})$ parameter regions distinguish from those of the Slovenian texts shown in Figure 8.6b, mostly for private letters, poems and prose.

Plotting the parameters of the Singh-Poisson model versus each other lead to a good discrimination of three Slovenian text groups, as can be seen in Figure 8.8a. Although some Russian texts are located in clearly defined areas there are many overlappings. However, some general tendency of journalistic texts to build a separate category, can still be observed (cf. Figure 8.8b). One possible explanation can be found in the fact that the journalistic texts originate from a Russian quality newspaper being as such free of colloquial speech, for which the use of nouns and compound words is quite typical. Furthermore, these types of words are featured by longer word forms, and hence location of the journalistic texts in the upper area of Figure 8.8b seems plausible. Although determined by the use of poetic meters, the Russian poems can not be so clearly distinguished from letters and prose texts, as the Slovenian do. This quite surprising phenomenon happens possibly due to higher heterogeneity of the Russian poems compared to Slovenian ones.



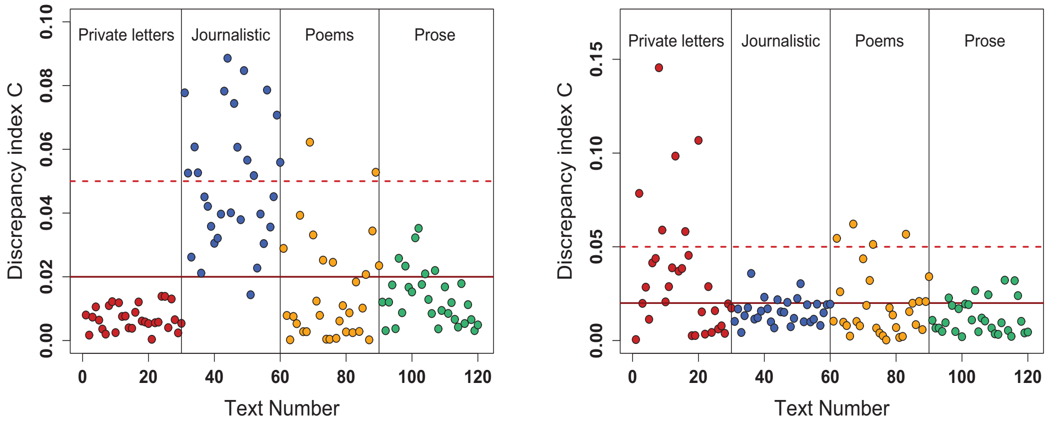
(a) Slovenian

(b) Russian

Figure 8.8: Text types discrimination by (α, θ) parameter range of Singh-Poisson model

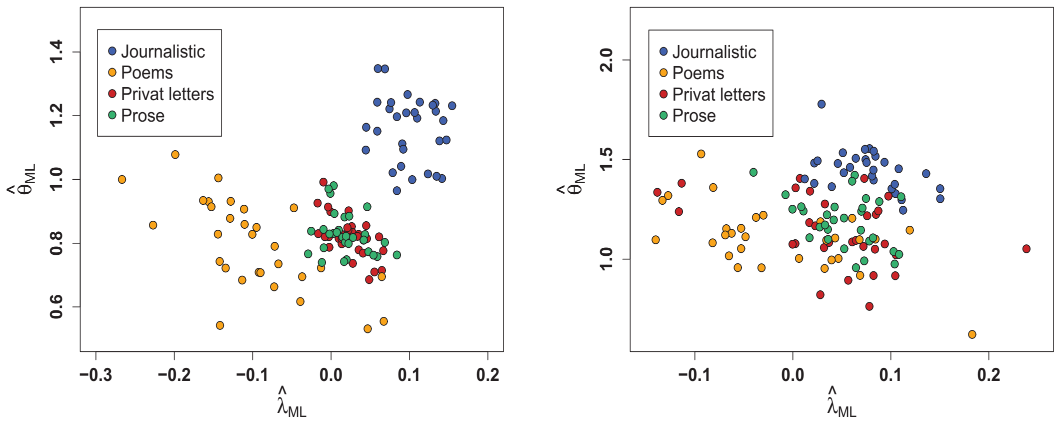
Furthermore, Figure 8.9a shows that the generalized Poisson model is not appropriate for almost half (46.67%) of the Slovenian journalistic texts. However, the existence of three different parameter patterns, namely for journalistic, poems and joint group of letters and prose texts, is evident from Figure 8.10a, regardless of the fact that the model fit for journalistic texts is not satisfactory.

As compared to this, this model provides more or less good fits for Russian texts, with exception of Achmatova short letters no. 2, 8, 9, 13, 16, and 20 of 130, 63, 151, 48, 201, and 75 words, respectively and poems no. 62, 73 and 83 of length 157, 302, and 109, respectively, as evident from Figure 8.9b. Yet, discrimination of Russian texts is again not so obvious as for the Slovenian text material.



(a) Slovenian (b) Russian

Figure 8.9: Results of fitting Generalized Poisson model to texts under study



(a) Slovenian (b) Russian

Figure 8.10: Text types discrimination by (λ, θ) parameter range of Generalized Poisson model

8.5 Summary

The analyzed Slavic languages, Slovenian and Russian, although belonging to the same family of Indo-European languages, can be attributed from both real and genetic point of view to two different subgroups. Russian is the largest of the three living members (together with Ukrainian and Belorussian) of the *East Slavic* languages. It is also the most widely spoken of all Slavic languages. Throughout history

Russian showed, however, different development compared to the Slovenian. The first disparities in the level of development date from 7th to 8th century and affect all linguistic levels, in particular the phonology, the syllable structure and morphology. Slovenian, on the contrary, belongs to the Western subgroup of the *South Slavic* branch of the Slavic languages, just like Serbian or Croatian. This language, located on the periphery of the Slavic languages, had large lexical influence from other non-Slavic languages, such as German or Italian, and hence has developed phonologically and morphologically in a different direction than Russian. This could be also demonstrated on the basis of word length and associated quantities in the present work. Nevertheless, we proved that the word length frequency distribution of our sampled texts can be theoretically described and that one discrete probability model is sufficient to describe them. This model, know in the literature as the *Singh-Poisson model*, is a simple two-parametric generalization of the Poisson distribution with parameter θ. The new parameter α tunes the type of dispersion. It allows to model under-dispersion ($1 < \alpha \leq \alpha_{\text{max}}$), equi-dispersion (Poisson case $\alpha = 1$) and over-dispersion ($0 < \alpha < 1$). Therefore, the proposed model offers a unified approach for all cases of under-, equi and over-dispersion. An additional benefit is that the maximum likelihood estimation leads, in case of the Singh-Poisson distribution, to the same estimates as the method based on the sample mean and the first frequency class. For this reason, the calculation of maximum likelihood estimates is a very simple task. In a simulation study we demonstrated the usefulness of the parameter estimates under three data-driven dispersion scenarios. Finally, the Singh-Poisson model is applied to 120 Slovenian, as well as 120 Russian texts, and in all cases we obtained reasonable and stable estimates.

Appendix A

Text Sources

This appendix gives an overview of the Slovenian and Russian texts, which are considered to represent the text corpus of this study. The detailed reference for the private letters (text number 1 to 30), journalistic texts (text number 31 to 60), poems (text number 61 to 90), and prose texts (text number 91 to 120) are given in Table A.1 for Slovenian and Table A.7 for Russian texts under study. Tables A.2 and A.8 show frequency distributions for each Slovenian and Russian text, respectively, as well as their text lengths TL, and three crucial statistical measures, namely mean word length $\bar{x}$, sample variance s^2, and index of dispersion $d = s^2/(\bar{x} - 1)$. Subsequently, Tables A.3 to A.6 display parameter estimation results of our specified models for Slovenian text, whereas Tables A.9 to A.12 highlights results for the Russian texts under study.

A.1 Slovenian Texts

Table A.1: Sources of the Slovenian texts

Text No.	Author	Title	Year
1-24	Cankar, Ivan	Letter to Ana Lušinova	1898
25-29	Cankar, Ivan	Letter to Ana Lušinova	1902
30	Cankar, Ivan	Letter to Ana Lušinova	1904
31	Journal Delo	Baron	2001
32	Journal Delo	Čistilka za ustavno sodišče	2001
33	Journal Delo	Dete je žur	2001
34	Journal Delo	Duhovi in duhovi	2001
35	Journal Delo	Kaj je z napadi	2001
36	Journal Delo	Kaj vse potrebujemo One?	2001
37	Journal Delo	Kam z Jelinčičem?	2001
38	Journal Delo	Ko še regulatorju ni jasno	2001
39	Journal Delo	Kriza vpliva tudi na obrambno moč držav	2001
40	Journal Delo	Motivi	2001
41	Journal Delo	Mrtvaški marš	2001
42	Journal Delo	Ne bobnajmo v prazno	2001

Table A.1: Sources of the Slovenian texts

Text No.	Author	Title	Year
43	Journal Delo	Ne spreglejmo bistva	2001
44	Journal Delo	Obrezovanje po rusko	2001
45	Journal Delo	Odhod z odra	2001
46	Journal Delo	Odločila bo Sadamova modrost	2001
47	Journal Delo	Omnibus politika	2001
48	Journal Delo	Oni pa da niso packi	2001
49	Journal Delo	Po sili razmer na istem bregu	2001
50	Journal Delo	Presežek demokracije	2001
51	Journal Delo	Protizakonite samoumevnosti	2001
52	Journal Delo	S porcelanastimi skodelicami na glavi	2001
53	Journal Delo	Strah pred kroglo	2001
54	Journal Delo	Teroristični koncert	2001
55	Journal Delo	Treznitev z vodo	2001
56	Journal Delo	Tržaški župan zaprl vrata manjšini	2001
57	Journal Delo	Verjetno bo zakonodajo treba spremeniti, toda po legalni poti	2001
58	Journal Delo	Vpliv uniforme na rast književnikov	2001
59	Journal Delo	Vse je mogoče pokvariti	2001
60	Journal Delo	Za uvod	2001
61	Gregorčič, Simon	Čas	1888
62	Gregorčič, Simon	Človeka nikar!	1877
63	Gregorčič, Simon	Cvete, cvete pomlad	1901
64	Gregorčič, Simon	Domovini	1880
65	Gregorčič, Simon	Kako srčno sva se ljubila	1901
66	Gregorčič, Simon	Kesanje	1882
67	Gregorčič, Simon	Moj crni plašč	1879
68	Gregorčič, Simon	Na potujčeni zemlji	1880
69	Gregorčič, Simon	Na sveti večer	1882
70	Gregorčič, Simon	Njega ni!	1879
71	Gregorčič, Simon	O nevihti	1878
72	Gregorčič, Simon	Oj zbogom, ti planinski svet!	1879
73	Gregorčič, Simon	Oljki	1882
74	Gregorčič, Simon	Pogled v nedolžno oko	1882
75	Gregorčič, Simon	Pozabljenim	1881
76	Gregorčič, Simon	Pri zibelki	1882
77	Gregorčič, Simon	Primuli	1882
78	Gregorčič, Simon	Sam	1872
79	Gregorčič, Simon	Samostanski vratar	1882
80	Gregorčič, Simon	Siroti	1882
81	Gregorčič, Simon	Sveta odkletev	1882
82	Gregorčič, Simon	Ti veselo poj!	1879
83	Gregorčič, Simon	Tri lipe	1878
84	Gregorčič, Simon	Ujetega ptica tožba	1878
85	Gregorčič, Simon	Veseli pastir	1871
86	Gregorčič, Simon	Zaostali ptič	1876
87	Gregorčič, Simon	Zimski dan	1879
88	Gregorčič, Simon	Življenje ni praznik	1878
89	Vodnik, Valentin	Vršac	1806
90	Vodnik, Valentin	Ilirija Oživljena	1811

Table A.1: Sources of the Slovenian texts

Text No.	Author	Title	Year
91-108	Cankar, Ivan	Hlapec Jernej in njegova pravica, Ch. 1-18	1907
109-117	Cankar, Ivan	Hiša Marije pomočnice, Ch. 1-9	1904
118	Cankar, Ivan	Mimo življenja	1920
119	Cankar, Ivan	O prešcah	1920
120	Cankar, Ivan	Brez doma	1903

Table A.2: Slovenian texts - frequency distribution and characteristic statistical measures

Text	Frequency classes												
No.	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	TL	$\bar{x}$	s^2	d
1	384	280	157	43	8	2	0	0	0	874	1.875	0.920	1.051
2	456	407	190	54	9	3	1	0	0	1120	1.898	0.898	0.999
3	451	370	195	65	8	1	0	0	0	1090	1.910	0.905	0.994
4	553	370	210	61	12	2	0	0	0	1208	1.853	0.929	1.088
5	712	543	242	79	8	0	0	0	0	1584	1.818	0.815	0.996
6	582	436	220	59	16	2	0	0	0	1315	1.857	0.903	1.054
7	747	576	277	84	17	3	0	0	0	1704	1.860	0.897	1.043
8	595	415	238	61	10	3	1	0	0	1323	1.858	0.918	1.070
9	601	447	207	69	3	0	0	0	0	1327	1.814	0.803	0.987
10	390	292	129	50	14	2	0	0	0	877	1.873	0.985	1.128
11	390	240	92	38	3	0	0	0	0	763	1.721	0.789	1.095
12	511	443	225	73	5	3	0	0	0	1260	1.910	0.882	0.969
13	726	526	260	77	7	2	0	0	0	1598	1.823	0.838	1.019
14	504	422	162	55	7	0	0	0	0	1150	1.817	0.792	0.970
15	407	310	130	43	4	2	0	0	0	896	1.809	0.832	1.028
16	903	638	333	92	11	2	0	0	0	1979	1.826	0.847	1.026
17	759	495	268	73	10	0	0	1	0	1606	1.808	0.863	1.069
18	680	422	183	64	9	1	0	0	0	1359	1.751	0.834	1.109
19	581	378	176	45	9	0	0	0	0	1189	1.758	0.798	1.053
20	681	415	194	62	11	2	0	0	0	1365	1.764	0.864	1.130
21	570	457	199	59	11	3	0	0	0	1299	1.840	0.862	1.026
22	883	619	262	100	12	3	0	0	0	1879	1.801	0.866	1.081
23	512	400	186	68	7	3	0	0	0	1176	1.866	0.902	1.041
24	239	204	120	45	6	0	0	0	0	614	1.982	0.967	0.985
25	184	163	65	34	4	1	0	0	0	451	1.922	0.965	1.046
26	399	313	147	50	17	1	0	0	0	927	1.895	0.977	1.091
27	267	211	68	30	12	1	0	0	0	589	1.832	0.956	1.150
28	232	164	88	24	5	1	0	0	0	514	1.850	0.907	1.067
29	181	149	51	16	4	0	0	0	0	401	1.786	0.784	0.998
30	252	177	89	29	5	1	0	0	0	553	1.844	0.914	1.083
31	298	200	202	131	29	6	0	0	0	866	2.320	1.494	1.132
32	388	230	219	106	22	7	4	0	0	976	2.161	1.452	1.251
33	357	300	242	137	52	8	1	2	1	1100	2.337	1.605	1.200
34	304	189	202	88	29	4	1	0	0	817	2.223	1.431	1.170
35	192	133	129	75	40	9	1	0	0	579	2.428	1.806	1.264
36	218	157	107	53	19	3	0	0	0	557	2.115	1.346	1.208

Table A.2: Slovenian texts - frequency distribution and characteristic statistical measures

Text	Frequency classes												
No.	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	TL	$\bar{x}$	s^2	d
37	251	167	156	75	21	4	1	0	0	675	2.206	1.419	1.177
38	198	148	145	84	29	5	0	4	0	613	2.401	1.741	1.242
39	205	146	147	72	35	9	3	1	0	618	2.401	1.800	1.284
40	301	254	219	94	14	9	1	1	0	893	2.218	1.328	1.090
41	385	267	213	97	20	5	1	0	0	988	2.108	1.276	1.151
42	177	119	116	65	27	8	3	0	0	515	2.383	1.797	1.300
43	256	148	210	91	33	15	8	1	1	763	2.456	1.960	1.346
44	222	147	146	100	16	2	3	0	0	636	2.307	1.511	1.156
45	310	200	155	73	11	4	0	0	0	753	2.053	1.221	1.159
46	197	138	182	82	36	3	1	2	0	641	2.446	1.638	1.133
47	341	247	228	163	43	6	11	1	1	1041	2.417	1.811	1.278
48	235	160	110	70	14	11	0	0	0	600	2.168	1.519	1.300
49	168	114	137	86	22	6	0	0	0	533	2.433	1.584	1.105
50	203	150	119	92	23	7	1	0	0	595	2.339	1.625	1.213
51	273	239	161	72	23	1	1	0	0	770	2.143	1.241	1.086
52	257	174	136	98	35	7	1	0	0	708	2.301	1.659	1.275
53	442	321	249	113	32	7	2	0	0	1166	2.143	1.347	1.178
54	386	271	222	141	36	26	3	0	0	1085	2.318	1.732	1.314
55	265	224	188	114	29	8	2	1	0	831	2.344	1.549	1.152
56	188	113	134	75	23	5	0	1	0	539	2.354	1.631	1.204
57	149	95	75	41	14	2	1	0	0	377	2.167	1.491	1.277
58	107	81	88	37	11	3	1	0	0	328	2.320	1.472	1.115
59	148	103	108	57	32	3	0	0	0	451	2.404	1.659	1.182
60	232	169	148	86	15	2	1	0	0	653	2.224	1.355	1.107
61	71	62	34	2	0	0	0	0	0	169	1.805	0.634	0.788
62	113	66	33	7	3	0	0	0	0	222	1.743	0.843	1.135
63	48	40	10	1	0	0	0	0	0	99	1.636	0.499	0.784
64	75	48	22	5	0	0	0	0	0	150	1.713	0.703	0.985
65	55	42	18	3	0	0	0	0	0	118	1.737	0.657	0.891
66	72	63	42	4	0	0	0	0	0	181	1.878	0.707	0.805
67	92	44	16	6	0	0	0	0	0	158	1.595	0.676	1.136
68	105	63	21	3	0	0	0	0	0	192	1.594	0.557	0.937
69	55	37	24	1	0	0	0	0	0	117	1.752	0.654	0.869
70	42	55	8	1	0	0	0	0	0	106	1.698	0.422	0.605
71	100	84	35	2	0	0	0	0	0	221	1.724	0.573	0.792
72	97	50	12	6	0	0	0	0	0	165	1.558	0.614	1.101
73	264	226	125	9	2	0	0	0	0	626	1.816	0.675	0.827
74	68	47	14	2	0	0	0	0	0	131	1.618	0.546	0.882
75	65	49	15	2	0	0	0	0	0	131	1.649	0.553	0.852
76	45	48	21	0	0	0	0	0	0	114	1.789	0.539	0.683
77	59	49	14	1	0	0	0	0	0	123	1.650	0.508	0.781
78	24	23	10	1	0	0	0	0	0	58	1.793	0.623	0.786
79	99	91	45	6	0	0	0	0	0	241	1.826	0.670	0.811
80	47	34	10	3	0	0	0	0	0	94	1.670	0.632	0.943
81	99	69	23	1	0	0	0	0	0	192	1.615	0.510	0.830
82	105	63	11	0	0	0	0	0	0	179	1.475	0.374	0.788
83	145	117	59	6	0	0	0	0	0	327	1.774	0.648	0.838
84	116	89	32	3	1	0	0	0	0	241	1.689	0.607	0.881

Table A.2: Slovenian texts - frequency distribution and characteristic statistical measures

Text	Frequency classes											
No.	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	TL	$\bar{x}$	s^2	d
85	54	52	24	2	0	0	0	0	0	132	1.803	0.617	0.769
86	59	49	25	2	0	0	0	0	0	135	1.778	0.637	0.819
87	123	93	28	4	0	0	0	0	0	248	1.649	0.553	0.851
88	79	58	34	3	0	0	0	0	0	174	1.776	0.672	0.866
89	49	84	19	7	0	0	0	0	0	159	1.899	0.597	0.664
90	111	85	54	10	1	0	0	0	0	261	1.870	0.806	0.927
91	259	195	92	35	3	0	0	0	0	584	1.849	0.866	1.019
92	461	325	130	33	3	0	0	0	0	952	1.731	0.716	0.980
93	497	315	164	37	6	0	0	0	0	1019	1.763	0.792	1.037
94	372	249	131	32	1	0	0	0	0	785	1.778	0.767	0.986
95	373	280	119	19	3	1	0	0	0	795	1.745	0.704	0.946
96	427	237	151	50	10	0	0	0	0	875	1.833	0.965	1.159
97	430	306	157	46	7	0	0	0	0	946	1.831	0.854	1.028
98	646	449	270	52	4	0	0	0	0	1421	1.817	0.783	0.959
99	406	288	168	37	5	0	0	0	0	904	1.835	0.822	0.984
100	544	337	180	39	5	0	0	0	0	1105	1.755	0.778	1.030
101	433	269	165	49	1	0	0	0	0	917	1.818	0.843	1.031
102	547	339	213	71	2	0	0	0	0	1172	1.841	0.882	1.048
103	693	477	283	61	11	0	0	0	0	1525	1.833	0.838	1.006
104	454	265	156	48	7	0	0	0	0	930	1.805	0.893	1.109
105	613	423	230	69	9	0	0	0	0	1344	1.838	0.867	1.035
106	555	361	185	58	10	1	0	0	0	1170	1.812	0.887	1.092
107	566	339	188	67	6	0	0	0	0	1166	1.806	0.886	1.099
108	129	93	46	16	4	0	0	0	0	288	1.865	0.940	1.087
109	1088	950	491	197	21	3	1	0	0	2751	1.955	0.954	0.999
110	1060	878	520	209	23	1	1	0	0	2692	1.984	0.990	1.006
111	1353	1058	578	180	23	4	0	0	0	3196	1.897	0.904	1.008
112	1632	1104	580	174	18	2	0	0	0	3510	1.817	0.851	1.041
113	1877	1450	775	251	43	5	0	0	0	4401	1.898	0.927	1.032
114	1612	1232	583	180	25	3	0	0	0	3635	1.840	0.852	1.014
115	1259	900	568	218	37	5	0	0	0	2987	1.958	1.040	1.086
116	1415	1006	496	165	25	8	0	0	0	3115	1.845	0.915	1.083
117	1085	833	438	168	17	5	0	0	0	2546	1.906	0.945	1.044
118	522	464	243	93	15	1	0	0	0	1338	1.967	0.962	0.995
119	1908	1521	656	208	22	6	0	0	0	4321	1.827	0.824	0.996
120	428	313	131	58	14	2	0	0	0	946	1.862	0.985	1.143

Table A.3: Singh-Poisson model - estimation results for Slovenian texts

		1-displaced SP				size-biased SP
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{MM}} = \hat{\theta}_{\mathrm{ML}}$
1	1.051	0.946	0.925	0.904	0.969	0.875
2	0.999	1.002	0.897	1.001	0.898	0.898
3	0.994	1.008	0.903	0.953	0.955	0.910
4	1.088	0.907	0.941	0.864	0.988	0.853

Table A.3: Singh-Poisson model - estimation results for Slovenian texts

		1-displaced SP				**size-biased SP**
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{MM}} = \hat{\theta}_{\mathrm{ML}}$
5	0.996	1.006	0.813	0.959	0.853	0.818
6	1.054	0.941	0.910	0.919	0.932	0.857
7	1.043	0.953	0.902	0.933	0.922	0.860
8	1.070	0.926	0.927	0.889	0.965	0.858
9	0.987	1.017	0.800	0.952	0.855	0.814
10	1.128	0.873	1.000	0.886	0.986	0.873
11	1.095	0.885	0.815	0.864	0.834	0.721
12	0.969	1.036	0.879	0.986	0.923	0.910
13	1.019	0.978	0.841	0.928	0.887	0.823
14	0.970	1.039	0.786	1.019	0.801	0.817
15	1.028	0.968	0.836	0.955	0.847	0.809
16	1.026	0.971	0.851	0.914	0.903	0.826
17	1.069	0.922	0.876	0.875	0.923	0.808
18	1.109	0.874	0.860	0.854	0.880	0.751
19	1.053	0.935	0.810	0.896	0.846	0.758
20	1.130	0.855	0.894	0.837	0.913	0.764
21	1.026	0.971	0.865	0.967	0.869	0.840
22	1.081	0.909	0.882	0.898	0.893	0.801
23	1.041	0.956	0.907	0.934	0.928	0.866
24	0.985	1.018	0.965	0.945	1.039	0.982
25	1.046	0.955	0.966	0.957	0.964	0.922
26	1.091	0.908	0.986	0.909	0.985	0.895
27	1.150	0.849	0.980	0.916	0.908	0.832
28	1.067	0.929	0.915	0.894	0.951	0.850
29	0.998	1.006	0.781	1.025	0.767	0.786
30	1.083	0.913	0.925	0.886	0.954	0.844
31	1.132	0.910	1.451	0.820	1.610	1.320
32	1.251	0.823	1.410	0.777	1.495	1.161
33	1.200	0.870	1.536	0.854	1.566	1.337
34	1.170	0.879	1.391	0.803	1.523	1.223
35	1.264	0.845	1.690	0.805	1.775	1.428
36	1.208	0.844	1.320	0.818	1.363	1.115
37	1.177	0.873	1.381	0.812	1.485	1.206
38	1.242	0.854	1.641	0.831	1.687	1.401
39	1.284	0.832	1.684	0.814	1.722	1.401
40	1.090	0.932	1.307	0.888	1.372	1.218
41	1.151	0.881	1.258	0.827	1.341	1.108
42	1.300	0.823	1.680	0.797	1.735	1.383
43	1.346	0.809	1.800	0.789	1.845	1.456
44	1.156	0.894	1.461	0.815	1.603	1.307
45	1.159	0.870	1.211	0.807	1.304	1.053
46	1.133	0.917	1.577	0.846	1.710	1.446
47	1.278	0.837	1.694	0.816	1.736	1.417
48	1.300	0.797	1.466	0.786	1.486	1.168
49	1.105	0.933	1.537	0.835	1.717	1.433
50	1.213	0.864	1.551	0.818	1.638	1.339
51	1.086	0.931	1.227	0.895	1.277	1.143
52	1.275	0.826	1.574	0.788	1.650	1.301

Table A.3: Singh-Poisson model - estimation results for Slovenian texts

		1-displaced SP				**size-biased SP**
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{MM}} = \hat{\theta}_{\mathrm{ML}}$
53	1.178	0.866	1.320	0.831	1.376	1.143
54	1.314	0.808	1.631	0.796	1.655	1.318
55	1.152	0.899	1.495	0.863	1.558	1.344
56	1.204	0.870	1.556	0.797	1.700	1.354
57	1.277	0.810	1.441	0.779	1.499	1.167
58	1.115	0.922	1.432	0.858	1.539	1.320
59	1.182	0.887	1.583	0.820	1.712	1.404
60	1.107	0.921	1.329	0.841	1.455	1.224
61	0.788	1.368	0.588	1.157	0.696	0.805
62	1.135	0.852	0.873	0.830	0.896	0.743
63	0.784	1.542	0.413	1.451	0.439	0.636
64	0.985	1.031	0.692	0.941	0.758	0.713
65	0.891	1.188	0.621	1.077	0.684	0.737
66	0.805	1.293	0.679	1.085	0.809	0.878
67	1.136	0.822	0.723	0.789	0.754	0.595
68	0.937	1.128	0.526	1.047	0.567	0.594
69	0.869	1.226	0.614	1.007	0.747	0.752
70	0.605	2.348	0.297	2.345	0.298	0.698
71	0.792	1.413	0.512	1.232	0.587	0.724
72	1.101	0.855	0.652	0.873	0.638	0.558
73	0.827	1.272	0.642	1.112	0.734	0.816
74	0.882	1.252	0.494	1.176	0.526	0.618
75	0.852	1.313	0.494	1.226	0.529	0.649
76	0.683	1.692	0.467	1.417	0.557	0.789
77	0.781	1.530	0.425	1.401	0.464	0.650
78	0.786	1.403	0.565	1.242	0.638	0.793
79	0.811	1.304	0.633	1.151	0.718	0.826
80	0.943	1.111	0.603	1.085	0.618	0.670
81	0.830	1.395	0.441	1.237	0.497	0.615
82	0.788	1.835	0.259	1.673	0.284	0.475
83	0.838	1.271	0.609	1.106	0.699	0.774
84	0.881	1.216	0.566	1.154	0.597	0.689
85	0.769	1.419	0.566	1.239	0.648	0.803
86	0.819	1.317	0.590	1.135	0.685	0.778
87	0.851	1.307	0.497	1.226	0.529	0.649
88	0.866	1.218	0.637	1.035	0.749	0.776
89	0.664	1.608	0.559	1.636	0.550	0.899
90	0.927	1.097	0.793	0.972	0.895	0.870
91	1.019	0.980	0.867	0.928	0.915	0.849
92	0.980	1.030	0.710	0.983	0.744	0.731
93	1.037	0.955	0.799	0.889	0.859	0.763
94	0.986	1.021	0.763	0.924	0.842	0.778
95	0.946	1.080	0.689	1.034	0.720	0.745
96	1.159	0.841	0.990	0.780	1.068	0.833
97	1.028	0.969	0.858	0.913	0.910	0.831
98	0.959	1.054	0.775	0.938	0.871	0.817
99	0.984	1.020	0.819	0.929	0.899	0.835
100	1.030	0.962	0.784	0.884	0.853	0.755

Table A.3: Singh-Poisson model - estimation results for Slovenian texts

		1-displaced SP				**size-biased SP**
No.	d	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\theta}_{\text{MM}} = \hat{\theta}_{\text{ML}}$
101	1.031	0.964	0.848	0.860	0.951	0.818
102	1.048	0.947	0.888	0.847	0.993	0.841
103	1.006	0.994	0.838	0.910	0.915	0.833
104	1.109	0.882	0.913	0.816	0.987	0.805
105	1.035	0.961	0.872	0.895	0.937	0.838
106	1.092	0.899	0.903	0.861	0.943	0.812
107	1.099	0.892	0.904	0.825	0.977	0.806
108	1.087	0.912	0.948	0.886	0.976	0.865
109	0.999	1.002	0.954	0.958	0.997	0.955
110	1.006	0.994	0.989	0.927	1.061	0.984
111	1.008	0.992	0.904	0.935	0.959	0.897
112	1.041	0.953	0.858	0.892	0.916	0.817
113	1.032	0.965	0.930	0.921	0.974	0.898
114	1.014	0.984	0.854	0.945	0.888	0.840
115	1.086	0.918	1.044	0.862	1.112	0.958
116	1.083	0.911	0.927	0.890	0.950	0.845
117	1.044	0.955	0.949	0.911	0.995	0.906
118	0.995	1.006	0.961	0.962	1.006	0.967
119	0.996	1.005	0.823	0.979	0.845	0.827
120	1.143	0.858	1.004	0.873	0.987	0.862

Table A.4: Hyper-Poisson model - estimation results for Slovenian texts

		1-displaced HP						**size-biased HP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
1	1.051	0.925	0.864	1.369	1.084	1.256	1.019	1.095	0.920	1.917	1.176	1.234	0.962
2	0.999	1.060	0.926	0.988	0.891	0.997	0.897	1.086	0.924	0.962	0.885	0.993	0.896
3	0.994	0.627	0.729	1.115	0.978	0.974	0.895	0.411	0.724	1.205	0.980	0.957	0.895
4	1.088	1.010	0.898	1.624	1.196	1.497	1.123	1.449	1.014	2.681	1.389	3.500	1.628
5	0.996	0.657	0.662	1.100	0.874	0.981	0.808	0.453	0.658	1.192	0.879	0.966	0.807
6	1.054	1.037	0.899	1.330	1.042	1.261	1.002	1.299	0.963	1.867	1.137	1.271	0.954
7	1.043	0.995	0.877	1.267	1.011	1.201	0.973	1.175	0.925	1.689	1.083	1.252	0.948
8	1.070	1.113	0.942	1.456	1.112	1.364	1.058	1.511	1.032	2.187	1.240	2.717	1.394
9	0.987	0.537	0.600	1.102	0.870	0.940	0.781	0.240	0.590	1.162	0.865	0.901	0.782
10	1.128	1.476	1.154	1.699	1.262	1.669	1.245	2.572	1.390	3.207	1.585	3.296	1.610
11	1.095	1.097	0.801	1.672	1.051	1.570	0.999	1.738	0.941	3.008	1.282	3.407	1.384
12	0.969	0.642	0.725	0.984	0.901	0.876	0.836	0.363	0.701	0.893	0.874	0.750	0.822
13	1.019	0.754	0.720	1.222	0.945	1.090	0.872	0.686	0.737	1.495	0.979	1.114	0.862
14	0.970	0.712	0.673	0.923	0.773	0.881	0.750	0.456	0.650	0.783	0.747	0.716	0.726
15	1.028	0.993	0.818	1.167	0.900	1.126	0.878	1.104	0.847	1.427	0.942	1.216	0.879
16	1.026	0.745	0.722	1.275	0.977	1.126	0.894	0.699	0.745	1.624	1.023	1.138	0.873
17	1.069	1.064	0.867	1.499	1.074	1.382	1.009	1.444	0.955	2.295	1.205	2.943	1.386
18	1.109	1.230	0.897	1.754	1.131	1.665	1.083	2.091	1.079	3.309	1.416	3.921	1.579
19	1.053	0.917	0.745	1.398	0.963	1.286	0.904	1.118	0.804	2.039	1.063	1.221	0.833

Table A.4: Hyper-Poisson model - estimation results for Slovenian texts

		1-displaced HP						**size-biased HP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
20	1.130	1.386	0.987	1.903	1.221	1.830	1.180	2.593	1.240	3.880	1.600	4.918	1.883
21	1.026	1.050	0.874	1.130	0.913	1.114	0.904	1.187	0.903	1.344	0.950	1.238	0.918
22	1.081	1.150	0.904	1.489	1.061	1.423	1.025	1.674	1.017	2.408	1.231	2.590	1.279
23	1.041	0.960	0.867	1.264	1.016	1.190	0.974	1.107	0.911	1.671	1.086	1.244	0.953
24	0.985	0.464	0.708	1.130	1.063	0.934	0.942	0.149	0.696	1.204	1.056	0.896	0.943
25	1.046	0.999	0.943	1.225	1.055	1.198	1.040	1.178	0.992	1.621	1.135	1.372	1.055
26	1.091	1.226	1.045	1.486	1.172	1.444	1.148	1.819	1.177	2.408	1.362	2.513	1.391
27	1.150	1.959	1.334	1.673	1.195	1.734	1.233	3.813	1.702	3.511	1.611	2.791	1.402
28	1.067	1.015	0.887	1.441	1.095	1.347	1.040	1.327	0.967	2.157	1.220	1.268	0.950
29	0.998	0.983	0.776	0.949	0.758	0.992	0.781	0.955	0.772	0.899	0.754	0.972	0.777
30	1.083	1.073	0.915	1.526	1.133	1.437	1.082	1.519	1.022	2.443	1.302	2.877	1.426
31	1.132	0.418	1.063	2.113	2.083	1.864	1.886	0.554	1.195	3.590	2.427	1.368	1.519
32	1.251	1.537	1.583	2.833	2.309	3.066	2.405	3.526	2.180	6.705	3.359	1.467	1.424
33	1.200	1.554	1.768	2.164	2.139	2.109	2.086	2.812	2.143	4.206	2.712	1.561	1.642
34	1.170	0.804	1.208	2.288	2.067	2.202	1.978	1.397	1.437	4.315	2.563	1.404	1.440
35	1.264	1.091	1.636	2.868	2.725	2.938	2.724	2.375	2.110	6.250	3.744	1.588	1.783
36	1.208	1.268	1.365	2.420	1.996	2.344	1.933	2.606	1.763	5.296	2.737	1.458	1.353
37	1.177	0.948	1.273	2.273	2.035	2.185	1.950	1.695	1.533	4.375	2.557	1.423	1.432
38	1.242	1.815	2.017	2.449	2.410	2.512	2.425	3.530	2.543	5.052	3.180	1.597	1.747
39	1.284	1.749	2.003	2.830	2.662	2.959	2.710	3.819	2.679	6.493	3.793	1.625	1.776
40	1.090	1.100	1.322	1.556	1.594	1.434	1.506	1.482	1.434	2.321	1.762	1.358	1.388
41	1.151	1.004	1.189	2.069	1.779	1.940	1.682	1.697	1.410	3.862	2.195	1.385	1.300
42	1.300	1.604	1.905	3.067	2.789	3.288	2.884	3.741	2.624	7.346	4.112	1.610	1.754
43	1.346	2.461	2.526	3.249	3.016	3.862	3.358	5.884	3.659	8.273	4.675	1.672	1.885
44	1.156	0.718	1.234	2.198	2.121	2.047	1.988	1.145	1.425	3.904	2.536	1.407	1.529
45	1.159	0.928	1.097	2.199	1.780	2.082	1.690	1.648	1.331	4.282	2.253	1.364	1.233
46	1.133	0.895	1.462	1.938	2.123	1.773	1.982	1.291	1.623	3.137	2.419	1.421	1.678
47	1.278	1.904	2.109	2.749	2.631	2.890	2.688	4.022	2.785	6.141	3.675	1.625	1.791
48	1.300	1.862	1.794	3.188	2.531	3.328	2.584	4.932	2.728	8.810	4.167	1.534	1.476
49	1.105	0.317	1.092	1.959	2.122	1.648	1.877	0.298	1.184	3.068	2.370	1.336	1.618
50	1.213	0.955	1.435	2.504	2.363	2.416	2.272	1.851	1.766	5.039	3.063	1.515	1.632
51	1.086	0.886	1.127	1.556	1.506	1.408	1.406	1.141	1.225	2.370	1.683	1.353	1.303
52	1.275	1.186	1.561	3.051	2.655	3.173	2.685	2.823	2.127	7.216	3.875	1.545	1.624
53	1.178	1.246	1.369	2.171	1.887	2.086	1.817	2.289	1.676	4.299	2.419	1.441	1.368
54	1.314	1.851	1.980	3.181	2.763	3.360	2.838	4.605	2.874	8.257	4.336	1.608	1.684
55	1.152	1.114	1.496	1.942	2.000	1.811	1.896	1.765	1.712	3.382	2.377	1.486	1.601
56	1.204	0.907	1.417	2.506	2.386	2.545	2.361	1.698	1.718	4.866	3.017	1.471	1.627
57	1.277	1.550	1.611	3.052	2.449	3.254	2.530	3.862	2.323	7.787	3.780	1.500	1.453
58	1.115	0.866	1.308	1.793	1.872	1.625	1.741	1.192	1.441	2.843	2.114	1.385	1.519
59	1.182	0.607	1.279	2.355	2.352	2.209	2.216	1.050	1.498	4.363	2.888	1.476	1.675
60	1.107	0.579	1.050	1.876	1.807	1.625	1.627	0.725	1.159	3.037	2.057	1.332	1.390
61	0.788	∅	0.260	0.540	0.535	0.295	0.396	∅	0.215	∅	0.435	∅	∅
62	1.135	1.405	0.971	1.921	1.200	1.888	1.179	2.670	1.231	3.973	1.588	5.328	1.952
63	0.784	0.230	0.250	0.364	0.308	0.336	0.294	∅	0.198	∅	0.239	∅	∅
64	0.985	0.459	0.479	1.097	0.762	0.924	0.676	0.078	0.465	1.120	0.747	0.886	0.679
65	0.891	0.279	0.382	0.717	0.585	0.577	0.512	∅	0.336	0.243	0.511	0.097	0.465
66	0.805	∅	0.301	0.627	0.650	0.323	0.471	∅	0.254	0.027	0.541	∅	∅
67	1.136	1.270	0.744	2.183	1.094	2.132	1.068	2.753	1.026	5.202	1.592	8.328	2.306

Table A.4: Hyper-Poisson model - estimation results for Slovenian texts

		1-displaced HP						size-biased HP					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
68	0.937	0.455	0.366	0.816	0.510	0.711	0.463	∅	0.332	0.469	0.461	0.363	0.434
69	0.869	∅	0.258	0.787	0.637	0.441	0.456	∅	0.223	0.329	0.549	∅	∅
70	0.605	0.151	0.184	0.135	0.176	0.157	0.189	∅	0.135	∅	0.125	∅	∅
71	0.792	0.086	0.251	0.486	0.441	0.316	0.349	∅	0.205	∅	0.354	∅	∅
72	1.101	1.427	0.744	1.666	0.831	1.670	0.834	2.639	0.936	3.353	1.092	3.230	1.066
73	0.827	0.145	0.354	0.614	0.590	0.396	0.467	∅	0.301	0.018	0.494	∅	∅
74	0.882	0.374	0.331	0.602	0.426	0.545	0.399	∅	0.284	0.015	0.360	∅	∅
75	0.852	0.308	0.316	0.533	0.413	0.471	0.382	∅	0.265	∅	0.341	∅	∅
76	0.683	∅	0.177	0.325	0.378	0.148	0.274	∅	0.137	∅	0.287	∅	∅
77	0.781	0.177	0.237	0.385	0.329	0.316	0.295	∅	0.189	∅	0.256	∅	∅
78	0.786	0.095	0.292	0.463	0.476	0.341	0.407	∅	0.239	∅	0.381	∅	∅
79	0.811	0.108	0.335	0.562	0.565	0.375	0.458	∅	0.282	∅	0.468	∅	∅
80	0.943	0.586	0.477	0.766	0.553	0.768	0.554	0.140	0.438	0.397	0.504	0.417	0.510
81	0.830	0.168	0.232	0.514	0.378	0.378	0.313	∅	0.190	∅	0.306	∅	∅
82	0.788	0.151	0.132	0.309	0.189	0.255	0.167	∅	0.100	∅	0.142	∅	∅
83	0.838	0.110	0.314	0.634	0.568	0.411	0.446	∅	0.267	0.053	0.477	∅	∅
84	0.881	0.472	0.423	0.617	0.489	0.562	0.462	∅	0.368	0.050	0.417	0.003	0.403
85	0.769	0.056	0.277	0.460	0.481	0.293	0.385	∅	0.226	∅	0.384	∅	∅
86	0.819	0.065	0.287	0.584	0.541	0.368	0.422	∅	0.240	∅	0.447	∅	∅
87	0.851	0.324	0.323	0.535	0.414	0.469	0.382	∅	0.272	∅	0.342	∅	∅
88	0.866	0.071	0.309	0.749	0.637	0.464	0.483	∅	0.268	0.271	0.549	∅	∅
89	0.664	0.268	0.392	0.216	0.358	0.272	0.396	∅	0.308	∅	0.268	∅	∅
90	0.927	0.262	0.491	0.941	0.835	0.689	0.691	∅	0.454	0.708	0.772	0.446	0.680
91	1.019	0.654	0.698	1.236	0.982	1.091	0.900	0.527	0.714	1.531	1.021	1.118	0.890
92	0.980	0.657	0.577	1.007	0.735	0.908	0.684	0.387	0.561	0.954	0.718	0.815	0.676
93	1.037	0.729	0.662	1.372	0.957	1.202	0.867	0.721	0.696	1.884	1.025	1.160	0.818
94	0.986	0.434	0.524	1.154	0.861	0.929	0.741	0.071	0.514	1.248	0.854	0.904	0.747
95	0.946	0.684	0.586	0.845	0.662	0.777	0.626	0.331	0.548	0.560	0.613	0.498	0.594
96	1.159	1.044	0.925	2.322	1.528	2.253	1.475	2.108	1.205	5.098	2.093	1.295	0.967
97	1.028	0.707	0.710	1.291	0.991	1.137	0.906	0.644	0.734	1.670	1.043	1.147	0.882
98	0.959	0.337	0.500	1.071	0.856	0.807	0.712	∅	0.477	1.008	0.819	0.703	0.719
99	0.984	0.477	0.590	1.144	0.916	0.926	0.794	0.145	0.578	1.216	0.904	0.897	0.800
100	1.030	0.652	0.618	1.364	0.942	1.168	0.840	0.566	0.644	1.831	0.999	1.133	0.800
101	1.031	0.428	0.574	1.463	1.068	1.180	0.913	0.208	0.597	2.013	1.133	1.121	0.862
102	1.048	0.456	0.614	1.568	1.152	1.287	0.995	0.317	0.653	2.305	1.255	1.164	0.904
103	1.006	0.561	0.636	1.237	0.964	1.028	0.848	0.338	0.640	1.464	0.980	1.032	0.844
104	1.109	0.868	0.794	1.903	1.277	1.740	1.184	1.360	0.940	3.513	1.566	1.257	0.911
105	1.035	0.663	0.700	1.366	1.040	1.182	0.937	0.604	0.731	1.850	1.108	1.161	0.896
106	1.092	1.031	0.867	1.653	1.159	1.532	1.092	1.534	0.993	2.814	1.368	1.275	0.918
107	1.099	0.798	0.760	1.832	1.242	1.648	1.139	1.166	0.882	3.266	1.494	1.250	0.906
108	1.087	1.010	0.908	1.548	1.169	1.456	1.116	1.419	1.016	2.508	1.351	1.308	0.984
109	0.999	0.681	0.799	1.118	1.027	0.994	0.952	0.512	0.797	1.229	1.036	0.990	0.952
110	1.006	0.553	0.765	1.225	1.122	1.028	1.000	0.345	0.769	1.444	1.143	1.037	0.998
111	1.008	0.674	0.746	1.188	1.006	1.036	0.918	0.526	0.752	1.383	1.026	1.051	0.915
112	1.041	0.730	0.714	1.387	1.027	1.216	0.932	0.740	0.753	1.931	1.107	1.179	0.881
113	1.032	0.789	0.813	1.284	1.062	1.152	0.985	0.796	0.844	1.671	1.122	1.179	0.963
114	1.014	0.781	0.746	1.170	0.935	1.064	0.875	0.713	0.759	1.380	0.962	1.098	0.873
115	1.086	0.809	0.907	1.652	1.343	1.470	1.230	1.058	1.005	2.655	1.533	3.729	1.870

Table A.4: Hyper-Poisson model - estimation results for Slovenian texts

		1-displaced HP						**size-biased HP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
116	1.083	1.136	0.945	1.512	1.127	1.431	1.081	1.639	1.060	2.427	1.298	2.779	1.398
117	1.044	0.802	0.832	1.349	1.108	1.210	1.026	0.862	0.876	1.847	1.191	1.215	0.986
118	0.995	0.642	0.788	1.105	1.032	0.979	0.954	0.443	0.784	1.189	1.035	0.963	0.953
119	0.996	0.810	0.739	1.050	0.855	0.984	0.819	0.692	0.736	1.089	0.855	0.968	0.817
120	1.143	1.540	1.177	1.826	1.313	1.783	1.290	2.869	1.467	3.704	1.721	3.799	1.747

Table A.5: Generalized Poisson model - estimation results for Slovenian texts

		1-displaced GP						**size-biased GP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
1	1.051	0.025	0.854	0.026	0.852	0.060	0.822	0.024	0.805	0.026	0.800	0.058	0.704
2	0.999	0.000	0.899	-0.001	0.899	0.000	0.899	0.000	0.899	-0.001	0.900	0.000	0.899
3	0.994	-0.003	0.913	-0.004	0.914	0.030	0.882	-0.003	0.919	-0.004	0.921	0.030	0.822
4	1.088	0.041	0.818	0.046	0.814	0.084	0.781	0.040	0.736	0.045	0.724	0.080	0.617
5	0.996	-0.002	0.820	-0.003	0.820	0.023	0.800	-0.002	0.824	-0.003	0.826	0.022	0.755
6	1.054	0.026	0.835	0.027	0.834	0.049	0.815	0.026	0.783	0.027	0.780	0.048	0.719
7	1.043	0.021	0.842	0.022	0.841	0.041	0.825	0.021	0.800	0.022	0.797	0.040	0.744
8	1.070	0.033	0.829	0.035	0.828	0.069	0.799	0.033	0.764	0.034	0.758	0.066	0.665
9	0.987	-0.007	0.819	-0.008	0.820	0.027	0.792	-0.007	0.833	-0.008	0.836	0.026	0.739
10	1.128	0.058	0.822	0.061	0.820	0.072	0.810	0.056	0.708	0.059	0.700	0.069	0.669
11	1.095	0.044	0.689	0.049	0.686	0.069	0.671	0.043	0.602	0.047	0.590	0.066	0.537
12	0.969	-0.016	0.925	-0.017	0.926	0.009	0.902	-0.016	0.956	-0.017	0.960	0.009	0.885
13	1.019	0.009	0.815	0.010	0.815	0.041	0.789	0.009	0.797	0.010	0.795	0.040	0.708
14	0.970	-0.015	0.829	-0.016	0.830	-0.010	0.825	-0.015	0.860	-0.016	0.863	-0.010	0.846
15	1.028	0.014	0.798	0.014	0.798	0.025	0.789	0.014	0.771	0.014	0.771	0.024	0.740
16	1.026	0.013	0.815	0.014	0.814	0.050	0.785	0.012	0.790	0.014	0.787	0.048	0.687
17	1.069	0.033	0.781	0.035	0.779	0.072	0.750	0.032	0.716	0.035	0.709	0.069	0.610
18	1.109	0.051	0.713	0.055	0.710	0.078	0.692	0.049	0.614	0.054	0.601	0.074	0.541
19	1.053	0.026	0.738	0.028	0.737	0.055	0.716	0.025	0.688	0.027	0.682	0.053	0.609
20	1.130	0.059	0.719	0.065	0.715	0.090	0.695	0.057	0.603	0.062	0.588	0.085	0.522
21	1.026	0.013	0.829	0.013	0.829	0.019	0.824	0.013	0.804	0.013	0.804	0.019	0.785
22	1.081	0.038	0.771	0.041	0.769	0.058	0.755	0.037	0.696	0.040	0.688	0.056	0.642
23	1.041	0.020	0.849	0.021	0.848	0.040	0.832	0.020	0.810	0.021	0.807	0.039	0.752
24	0.985	-0.008	0.990	-0.010	0.992	0.039	0.944	-0.008	1.005	-0.010	1.011	0.038	0.866
25	1.046	0.022	0.902	0.022	0.902	0.028	0.897	0.022	0.857	0.022	0.857	0.028	0.841
26	1.091	0.043	0.857	0.045	0.855	0.058	0.843	0.042	0.773	0.044	0.767	0.057	0.728
27	1.150	0.067	0.776	0.067	0.776	0.049	0.791	0.065	0.644	0.064	0.647	0.048	0.695
28	1.067	0.032	0.823	0.034	0.822	0.064	0.795	0.031	0.760	0.033	0.755	0.062	0.670
29	0.998	-0.001	0.786	-0.002	0.787	-0.013	0.795	-0.001	0.788	-0.002	0.792	-0.013	0.821
30	1.083	0.039	0.812	0.041	0.810	0.069	0.786	0.038	0.735	0.040	0.728	0.066	0.651
31	1.132	0.060	1.241	0.074	1.222	0.192	1.067	0.059	1.121	0.073	1.072	0.176	0.696
32	1.251	0.106	1.038	0.124	1.017	0.205	0.922	0.101	0.830	0.118	0.772	0.186	0.531
33	1.200	0.087	1.221	0.096	1.209	0.158	1.125	0.084	1.048	0.093	1.018	0.148	0.816
34	1.170	0.076	1.130	0.091	1.112	0.191	0.989	0.073	0.981	0.088	0.930	0.175	0.621

Table A.5: Generalized Poisson model - estimation results for Slovenian texts

		1-displaced GP						**size-biased GP**					
No.	d	$\hat{\lambda}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\lambda}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\lambda}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\lambda}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\lambda}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\lambda}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$
35	1.264	0.111	1.270	0.133	1.239	0.227	1.104	0.106	1.051	0.128	0.972	0.207	0.665
36	1.208	0.090	1.015	0.104	1.000	0.159	0.938	0.086	0.838	0.099	0.795	0.147	0.633
37	1.177	0.078	1.112	0.092	1.095	0.180	0.989	0.076	0.957	0.089	0.911	0.165	0.643
38	1.242	0.103	1.257	0.113	1.243	0.194	1.130	0.099	1.054	0.109	1.017	0.179	0.754
39	1.284	0.118	1.236	0.134	1.214	0.213	1.103	0.113	1.004	0.128	0.947	0.195	0.693
40	1.090	0.042	1.167	0.045	1.164	0.107	1.087	0.041	1.083	0.044	1.075	0.102	0.877
41	1.151	0.068	1.033	0.079	1.021	0.150	0.942	0.066	0.899	0.076	0.865	0.139	0.654
42	1.300	0.123	1.213	0.143	1.185	0.227	1.068	0.117	0.970	0.137	0.898	0.207	0.630
43	1.346	0.138	1.255	0.154	1.231	0.250	1.092	0.131	0.983	0.147	0.923	0.226	0.611
44	1.156	0.070	1.215	0.084	1.197	0.194	1.053	0.068	1.077	0.082	1.029	0.179	0.677
45	1.159	0.071	0.978	0.084	0.965	0.157	0.887	0.069	0.838	0.081	0.798	0.145	0.586
46	1.133	0.060	1.359	0.069	1.347	0.184	1.180	0.059	1.239	0.067	1.210	0.171	0.821
47	1.278	0.115	1.253	0.130	1.233	0.212	1.116	0.111	1.025	0.124	0.974	0.195	0.705
48	1.300	0.123	1.025	0.142	1.003	0.198	0.937	0.117	0.784	0.134	0.723	0.180	0.559
49	1.105	0.049	1.363	0.060	1.348	0.195	1.155	0.048	1.266	0.059	1.228	0.180	0.776
50	1.213	0.092	1.216	0.110	1.192	0.197	1.075	0.089	1.034	0.106	0.973	0.181	0.694
51	1.086	0.040	1.097	0.044	1.092	0.093	1.037	0.040	1.017	0.044	1.004	0.089	0.855
52	1.275	0.114	1.152	0.138	1.121	0.221	1.013	0.109	0.926	0.132	0.844	0.200	0.590
53	1.178	0.079	1.053	0.089	1.041	0.152	0.970	0.076	0.898	0.086	0.864	0.141	0.677
54	1.314	0.128	1.150	0.147	1.124	0.216	1.033	0.121	0.899	0.140	0.831	0.196	0.619
55	1.152	0.068	1.252	0.076	1.241	0.150	1.143	0.067	1.116	0.075	1.089	0.141	0.851
56	1.204	0.089	1.234	0.107	1.210	0.222	1.053	0.086	1.059	0.103	0.996	0.202	0.626
57	1.277	0.115	1.033	0.135	1.010	0.205	0.928	0.109	0.807	0.128	0.743	0.185	0.538
58	1.115	0.053	1.250	0.059	1.242	0.151	1.120	0.052	1.145	0.058	1.125	0.142	0.825
59	1.182	0.080	1.291	0.098	1.266	0.206	1.114	0.078	1.132	0.095	1.070	0.189	0.715
60	1.107	0.050	1.163	0.059	1.151	0.154	1.035	0.049	1.064	0.058	1.034	0.144	0.736
61	0.788	-0.126	0.906	-0.156	0.930	-0.078	0.867	-0.136	1.171	-0.168	1.252	-0.081	1.027
62	1.135	0.061	0.698	0.065	0.695	0.091	0.675	0.059	0.578	0.062	0.568	0.086	0.500
63	0.784	-0.129	0.719	-0.134	0.722	-0.138	0.724	-0.143	0.995	-0.150	1.012	-0.153	1.019
64	0.985	-0.008	0.719	-0.013	0.722	0.028	0.693	-0.008	0.734	-0.013	0.748	0.028	0.637
65	0.891	-0.059	0.781	-0.072	0.790	-0.035	0.763	-0.062	0.903	-0.075	0.937	-0.036	0.835
66	0.805	-0.114	0.979	-0.144	1.005	-0.049	0.922	-0.122	1.216	-0.153	1.298	-0.051	1.022
67	1.136	0.062	0.558	0.067	0.555	0.091	0.541	0.059	0.439	0.064	0.424	0.084	0.368
68	0.937	-0.033	0.613	-0.039	0.617	-0.016	0.604	-0.034	0.680	-0.040	0.697	-0.017	0.637
69	0.869	-0.073	0.807	-0.100	0.828	-0.004	0.755	-0.076	0.957	-0.105	1.032	-0.004	0.762
70	0.605	-0.286	0.898	-0.227	0.857	-0.326	0.926	-0.345	1.540	-0.279	1.391	-0.398	1.659
71	0.792	-0.124	0.813	-0.144	0.828	-0.095	0.793	-0.134	1.074	-0.157	1.130	-0.102	0.991
72	1.101	0.047	0.531	0.047	0.532	0.047	0.531	0.045	0.440	0.045	0.441	0.045	0.440
73	0.827	-0.100	0.898	-0.111	0.907	-0.058	0.863	-0.106	1.104	-0.118	1.136	-0.060	0.981
74	0.882	-0.065	0.658	-0.073	0.663	-0.060	0.656	-0.068	0.792	-0.077	0.814	-0.063	0.781
75	0.852	-0.084	0.703	-0.092	0.709	-0.080	0.701	-0.089	0.878	-0.099	0.901	-0.085	0.868
76	0.683	-0.210	0.955	-0.267	1.000	-0.177	0.930	-0.238	1.407	-0.298	1.553	-0.197	1.306
77	0.781	-0.132	0.736	-0.142	0.743	-0.130	0.735	-0.146	1.018	-0.158	1.047	-0.143	1.011
78	0.786	-0.128	0.895	-0.153	0.914	-0.113	0.882	-0.139	1.164	-0.166	1.233	-0.120	1.117
79	0.811	-0.111	0.917	-0.128	0.931	-0.077	0.890	-0.118	1.147	-0.136	1.195	-0.081	1.049
80	0.943	-0.030	0.690	-0.037	0.695	-0.034	0.693	-0.030	0.751	-0.038	0.770	-0.035	0.763
81	0.830	-0.097	0.674	-0.113	0.684	-0.078	0.662	-0.105	0.880	-0.123	0.922	-0.083	0.825
82	0.788	-0.126	0.535	-0.142	0.542	-0.123	0.533	-0.145	0.816	-0.162	0.854	-0.141	0.806

Table A.5: Generalized Poisson model - estimation results for Slovenian texts

		1-displaced GP						**size-biased GP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
83	0.838	-0.093	0.845	-0.110	0.859	-0.051	0.813	-0.098	1.038	-0.117	1.086	-0.053	0.917
84	0.881	-0.065	0.734	-0.067	0.735	-0.062	0.731	-0.068	0.869	-0.071	0.874	-0.064	0.858
85	0.769	-0.140	0.916	-0.163	0.934	-0.113	0.894	-0.153	1.211	-0.178	1.275	-0.121	1.129
86	0.819	-0.105	0.860	-0.129	0.878	-0.064	0.828	-0.112	1.079	-0.138	1.143	-0.067	0.959
87	0.851	-0.084	0.704	-0.090	0.708	-0.080	0.701	-0.090	0.879	-0.096	0.896	-0.085	0.868
88	0.866	-0.074	0.834	-0.095	0.850	-0.018	0.790	-0.078	0.987	-0.100	1.043	-0.018	0.825
89	0.664	-0.227	1.104	-0.199	1.078	-0.309	1.177	-0.255	1.586	-0.223	1.506	-0.355	1.836
90	0.927	-0.039	0.903	-0.047	0.911	0.017	0.855	-0.040	0.982	-0.048	1.007	0.017	0.821
91	1.019	0.010	0.841	0.010	0.841	0.043	0.813	0.009	0.822	0.010	0.822	0.042	0.729
92	0.980	-0.010	0.739	-0.012	0.740	0.008	0.725	-0.010	0.759	-0.012	0.763	0.008	0.709
93	1.037	0.018	0.750	0.020	0.748	0.060	0.718	0.018	0.714	0.020	0.709	0.057	0.602
94	0.986	-0.007	0.784	-0.009	0.786	0.041	0.747	-0.007	0.799	-0.009	0.804	0.039	0.667
95	0.946	-0.028	0.766	-0.029	0.766	-0.016	0.757	-0.029	0.823	-0.030	0.826	-0.016	0.790
96	1.159	0.071	0.774	0.084	0.763	0.139	0.717	0.068	0.636	0.081	0.598	0.127	0.454
97	1.028	0.014	0.820	0.015	0.819	0.051	0.788	0.013	0.793	0.014	0.790	0.049	0.688
98	0.959	-0.021	0.834	-0.025	0.838	0.035	0.788	-0.022	0.877	-0.026	0.889	0.034	0.719
99	0.984	-0.008	0.842	-0.010	0.843	0.042	0.800	-0.008	0.858	-0.010	0.863	0.041	0.719
100	1.030	0.015	0.744	0.017	0.742	0.061	0.709	0.015	0.714	0.016	0.709	0.059	0.590
101	1.031	0.015	0.805	0.018	0.803	0.083	0.750	0.015	0.775	0.018	0.767	0.078	0.590
102	1.048	0.023	0.822	0.028	0.818	0.094	0.762	0.023	0.776	0.028	0.762	0.089	0.580
103	1.006	0.003	0.830	0.003	0.830	0.053	0.789	0.003	0.825	0.003	0.825	0.051	0.685
104	1.109	0.050	0.765	0.059	0.758	0.110	0.717	0.049	0.666	0.057	0.642	0.102	0.507
105	1.035	0.017	0.823	0.019	0.822	0.063	0.785	0.017	0.789	0.019	0.784	0.061	0.662
106	1.092	0.043	0.777	0.048	0.773	0.081	0.746	0.042	0.692	0.046	0.679	0.077	0.588
107	1.099	0.046	0.769	0.054	0.763	0.103	0.723	0.045	0.678	0.053	0.656	0.097	0.524
108	1.087	0.041	0.829	0.043	0.827	0.071	0.803	0.040	0.749	0.042	0.742	0.068	0.664
109	0.999	-0.001	0.956	-0.001	0.956	0.029	0.928	-0.001	0.957	-0.001	0.958	0.029	0.870
110	1.006	0.003	0.981	0.003	0.980	0.052	0.932	0.003	0.975	0.003	0.974	0.051	0.829
111	1.008	0.004	0.893	0.004	0.893	0.041	0.860	0.004	0.885	0.004	0.885	0.040	0.778
112	1.041	0.020	0.801	0.022	0.799	0.063	0.766	0.020	0.761	0.022	0.754	0.060	0.643
113	1.032	0.016	0.883	0.017	0.882	0.051	0.852	0.016	0.852	0.017	0.848	0.049	0.753
114	1.014	0.007	0.834	0.007	0.834	0.032	0.813	0.007	0.820	0.007	0.819	0.031	0.750
115	1.086	0.040	0.920	0.046	0.914	0.099	0.864	0.039	0.840	0.045	0.823	0.093	0.672
116	1.083	0.039	0.812	0.042	0.810	0.066	0.789	0.038	0.736	0.041	0.728	0.064	0.659
117	1.044	0.021	0.887	0.023	0.885	0.058	0.853	0.021	0.845	0.023	0.839	0.056	0.739
118	0.995	-0.003	0.970	-0.003	0.970	0.027	0.941	-0.003	0.975	-0.003	0.976	0.026	0.888
119	0.996	-0.002	0.829	-0.002	0.829	0.012	0.817	-0.002	0.832	-0.002	0.833	0.012	0.794
120	1.143	0.065	0.806	0.069	0.802	0.079	0.793	0.062	0.679	0.066	0.668	0.076	0.639

Table A.6: Cohen-Poisson model - estimation results for Slovenian texts

		1-displaced CP						**size-biased CP**			
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\alpha}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$
1	1.051	0.068	0.900	0.119	0.919	0.104	0.914	∅	0.925	∅	0.969
2	0.999	∅	0.897	0.004	0.900	∅	0.898	0.004	0.897	0.002	0.898
3	0.994	∅	0.907	0.067	0.934	0.051	0.929	0.019	0.903	∅	0.955
4	1.088	0.117	0.896	0.159	0.912	0.149	0.908	∅	0.941	∅	0.988
5	0.996	∅	0.816	0.050	0.836	0.041	0.833	0.014	0.813	∅	0.853
6	1.054	0.072	0.883	0.090	0.890	0.085	0.888	∅	0.910	∅	0.932
7	1.043	0.058	0.881	0.074	0.887	0.071	0.886	∅	0.902	∅	0.922
8	1.070	0.093	0.892	0.136	0.907	0.120	0.902	∅	0.927	∅	0.965
9	0.987	∅	0.807	0.064	0.837	0.048	0.831	0.038	0.800	∅	0.855
10	1.128	0.167	0.935	0.102	0.911	0.125	0.919	∅	1.000	∅	0.986
11	1.095	0.128	0.766	0.120	0.764	0.133	0.768	∅	0.815	∅	0.834
12	0.969	∅	0.894	0.032	0.922	0.014	0.916	0.083	0.879	∅	0.923
13	1.019	0.025	0.832	0.088	0.855	0.074	0.850	∅	0.841	∅	0.887
14	0.970	∅	0.801	∅	0.810	∅	0.810	0.082	0.786	0.042	0.801
15	1.028	0.037	0.822	0.044	0.825	0.045	0.825	∅	0.836	∅	0.847
16	1.026	0.034	0.838	0.107	0.864	0.089	0.858	∅	0.851	∅	0.903
17	1.069	0.092	0.841	0.147	0.861	0.130	0.855	∅	0.876	∅	0.923
18	1.109	0.146	0.804	0.139	0.802	0.147	0.804	∅	0.860	∅	0.880
19	1.053	0.072	0.784	0.112	0.798	0.103	0.795	∅	0.810	∅	0.846
20	1.130	0.172	0.826	0.161	0.822	0.167	0.825	∅	0.894	∅	0.913
21	1.026	0.035	0.852	0.033	0.852	0.034	0.852	∅	0.865	∅	0.869
22	1.081	0.108	0.841	0.095	0.836	0.105	0.839	∅	0.882	∅	0.893
23	1.041	0.054	0.886	0.070	0.892	0.070	0.892	∅	0.907	∅	0.928
24	0.985	∅	0.974	0.085	1.013	0.064	1.005	0.045	0.965	∅	1.039
25	1.046	0.059	0.944	0.026	0.932	0.047	0.940	∅	0.966	∅	0.964
26	1.091	0.120	0.939	0.086	0.927	0.100	0.932	∅	0.986	∅	0.985
27	1.150	0.193	0.903	0.036	0.845	0.087	0.864	∅	0.980	∅	0.908
28	1.067	0.088	0.882	0.123	0.895	0.113	0.892	∅	0.915	∅	0.951
29	0.998	∅	0.783	∅	0.773	∅	0.777	0.013	0.781	0.052	0.767
30	1.083	0.108	0.884	0.124	0.890	0.122	0.889	∅	0.925	∅	0.954
31	1.132	0.184	1.384	0.317	1.427	0.301	1.423	∅	1.451	∅	1.610
32	1.251	0.334	1.280	0.337	1.281	0.336	1.281	∅	1.410	∅	1.495
33	1.200	0.281	1.433	0.221	1.414	0.241	1.420	∅	1.536	∅	1.566
34	1.170	0.231	1.304	0.330	1.337	0.306	1.331	∅	1.391	∅	1.523
35	1.264	0.382	1.554	0.325	1.539	0.361	1.547	∅	1.690	∅	1.775
36	1.208	0.273	1.213	0.233	1.200	0.256	1.207	∅	1.320	∅	1.363
37	1.177	0.238	1.290	0.295	1.310	0.286	1.307	∅	1.381	∅	1.485
38	1.242	0.346	1.517	0.287	1.498	0.300	1.502	∅	1.641	∅	1.687
39	1.284	0.407	1.536	0.308	1.506	0.334	1.513	∅	1.684	∅	1.722
40	1.090	0.122	1.262	0.186	1.283	0.163	1.276	∅	1.307	∅	1.372
41	1.151	0.201	1.181	0.249	1.198	0.241	1.195	∅	1.258	∅	1.341
42	1.300	0.426	1.524	0.329	1.496	0.363	1.504	∅	1.680	∅	1.735
43	1.346	0.508	1.619	0.403	1.587	0.404	1.587	∅	1.800	∅	1.845
44	1.156	0.216	1.382	0.320	1.415	0.306	1.412	∅	1.461	∅	1.603
45	1.159	0.208	1.129	0.267	1.150	0.259	1.147	∅	1.211	∅	1.304
46	1.133	0.192	1.510	0.322	1.548	0.281	1.539	∅	1.577	∅	1.710
47	1.278	0.402	1.549	0.304	1.520	0.333	1.527	∅	1.694	∅	1.736

Table A.6: Cohen-Poisson model - estimation results for Slovenian texts

		1-displaced CP						**size-biased CP**			
No.	d	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\alpha}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$
48	1.300	0.398	1.309	0.276	1.270	0.322	1.283	∅	1.466	∅	1.486
49	1.105	0.151	1.484	0.333	1.540	0.300	1.533	∅	1.537	∅	1.717
50	1.213	0.298	1.441	0.279	1.437	0.309	1.445	∅	1.551	∅	1.638
51	1.086	0.114	1.184	0.142	1.194	0.143	1.195	∅	1.227	∅	1.277
52	1.275	0.381	1.431	0.310	1.411	0.356	1.423	∅	1.574	∅	1.650
53	1.178	0.237	1.229	0.237	1.229	0.241	1.230	∅	1.320	∅	1.376
54	1.314	0.438	1.466	0.298	1.424	0.345	1.436	∅	1.631	∅	1.655
55	1.152	0.214	1.418	0.219	1.420	0.226	1.422	∅	1.495	∅	1.558
56	1.204	0.287	1.452	0.371	1.478	0.355	1.474	∅	1.556	∅	1.700
57	1.277	0.366	1.297	0.306	1.278	0.334	1.286	∅	1.441	∅	1.499
58	1.115	0.157	1.375	0.263	1.408	0.230	1.400	∅	1.432	∅	1.539
59	1.182	0.259	1.491	0.318	1.510	0.322	1.511	∅	1.583	∅	1.712
60	1.107	0.145	1.275	0.258	1.313	0.241	1.309	∅	1.329	∅	1.455
61	0.788	∅	0.688	∅	0.784	∅	0.755	0.484	0.588	0.272	0.696
62	1.135	0.173	0.805	0.174	0.806	0.172	0.805	∅	0.873	∅	0.896
63	0.784	∅	0.512	∅	0.546	∅	0.536	0.531	0.413	0.482	0.439
64	0.985	∅	0.702	0.084	0.743	0.055	0.733	0.061	0.692	∅	0.758
65	0.891	∅	0.676	∅	0.728	∅	0.713	0.294	0.621	0.142	0.684
66	0.805	∅	0.772	0.003	0.879	∅	0.848	0.447	0.679	0.177	0.809
67	1.136	0.179	0.656	0.185	0.658	0.193	0.661	∅	0.723	∅	0.754
68	0.937	∅	0.559	∅	0.590	∅	0.582	0.192	0.526	0.078	0.567
69	0.869	∅	0.679	0.083	0.781	∅	0.750	0.340	0.614	0.015	0.747
70	0.605	∅	0.456	∅	0.473	∅	0.468	0.773	0.297	0.772	0.298
71	0.792	∅	0.609	∅	0.681	∅	0.659	0.488	0.512	0.339	0.587
72	1.101	0.138	0.603	0.076	0.582	0.107	0.593	∅	0.652	∅	0.638
73	0.827	∅	0.724	∅	0.804	∅	0.779	0.406	0.642	0.210	0.734
74	0.882	∅	0.553	∅	0.583	∅	0.574	0.330	0.494	0.254	0.526
75	0.852	∅	0.566	∅	0.602	∅	0.592	0.391	0.494	0.313	0.529
76	0.683	∅	0.607	∅	0.706	∅	0.675	0.652	0.467	0.514	0.557
77	0.781	∅	0.526	∅	0.571	∅	0.557	0.530	0.425	0.455	0.464
78	0.786	∅	0.670	∅	0.742	∅	0.720	0.506	0.565	0.369	0.638
79	0.811	∅	0.723	∅	0.798	∅	0.777	0.439	0.633	0.268	0.718
80	0.943	∅	0.636	∅	0.647	∅	0.646	0.183	0.603	0.145	0.618
81	0.830	∅	0.520	∅	0.574	∅	0.558	0.440	0.441	0.315	0.497
82	0.788	∅	0.351	∅	0.383	∅	0.372	0.589	0.259	0.534	0.284
83	0.838	∅	0.686	∅	0.763	∅	0.740	0.392	0.609	0.193	0.699
84	0.881	∅	0.625	∅	0.657	∅	0.646	0.313	0.566	0.242	0.597
85	0.769	∅	0.674	∅	0.754	∅	0.730	0.520	0.566	0.368	0.648
86	0.819	∅	0.678	∅	0.760	∅	0.736	0.435	0.590	0.236	0.685
87	0.851	∅	0.568	∅	0.601	∅	0.592	0.386	0.497	0.313	0.529
88	0.866	∅	0.703	0.041	0.790	∅	0.764	0.338	0.637	0.072	0.749
89	0.664	∅	0.709	∅	0.722	∅	0.722	0.661	0.559	0.673	0.550
90	0.927	∅	0.830	0.084	0.900	0.029	0.880	0.195	0.793	∅	0.895
91	1.019	0.024	0.858	0.082	0.879	0.075	0.877	∅	0.867	∅	0.915
92	0.980	∅	0.720	0.029	0.741	0.016	0.737	0.059	0.710	∅	0.744
93	1.037	0.050	0.781	0.134	0.811	0.111	0.803	∅	0.799	∅	0.859
94	0.986	∅	0.770	0.108	0.817	0.075	0.805	0.043	0.763	∅	0.842
95	0.946	∅	0.716	∅	0.742	∅	0.734	0.148	0.689	0.068	0.720

Table A.6: Cohen-Poisson model - estimation results for Slovenian texts

		1-displaced CP						size-biased CP			
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\alpha}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$
96	1.159	0.205	0.908	0.258	0.927	0.249	0.924	∅	0.990	∅	1.068
97	1.028	0.036	0.844	0.106	0.869	0.091	0.864	∅	0.858	∅	0.910
98	0.959	∅	0.796	0.111	0.857	0.063	0.840	0.111	0.775	∅	0.871
99	0.984	∅	0.827	0.111	0.875	0.074	0.862	0.045	0.819	∅	0.899
100	1.030	0.041	0.769	0.143	0.805	0.115	0.796	∅	0.784	∅	0.853
101	1.031	0.041	0.833	0.183	0.884	0.148	0.872	∅	0.848	∅	0.951
102	1.048	0.064	0.865	0.198	0.913	0.167	0.902	∅	0.888	∅	0.993
103	1.006	0.007	0.835	0.128	0.879	0.094	0.867	∅	0.838	∅	0.915
104	1.109	0.144	0.858	0.214	0.883	0.199	0.878	∅	0.913	∅	0.987
105	1.035	0.047	0.855	0.130	0.885	0.112	0.878	∅	0.872	∅	0.937
106	1.092	0.122	0.856	0.151	0.867	0.147	0.865	∅	0.903	∅	0.943
107	1.099	0.131	0.854	0.199	0.878	0.187	0.875	∅	0.904	∅	0.977
108	1.087	0.111	0.905	0.120	0.909	0.124	0.910	∅	0.948	∅	0.976
109	0.999	∅	0.954	0.056	0.976	0.047	0.973	0.005	0.954	∅	0.997
110	1.006	0.008	0.987	0.103	1.021	0.085	1.015	∅	0.989	∅	1.061
111	1.008	0.010	0.901	0.089	0.929	0.071	0.923	∅	0.904	∅	0.959
112	1.041	0.055	0.837	0.129	0.864	0.113	0.858	∅	0.858	∅	0.916
113	1.032	0.043	0.913	0.097	0.933	0.086	0.929	∅	0.930	∅	0.974
114	1.014	0.019	0.847	0.065	0.864	0.056	0.860	∅	0.854	∅	0.888
115	1.086	0.113	1.000	0.176	1.023	0.165	1.019	∅	1.044	∅	1.112
116	1.083	0.110	0.885	0.116	0.888	0.117	0.888	∅	0.927	∅	0.950
117	1.044	0.058	0.927	0.106	0.944	0.099	0.942	∅	0.949	∅	0.995
118	0.995	∅	0.964	0.052	0.986	0.043	0.983	0.016	0.961	∅	1.006
119	0.996	∅	0.825	0.026	0.837	0.021	0.835	0.011	0.823	∅	0.845
120	1.143	0.186	0.930	0.109	0.902	0.139	0.912	∅	1.004	∅	0.987

A.2 Russian Texts

Table A.7: Sources of the Russian texts

Text No.	Author	Title	Year
1	Achmatova, Anna Andreevna	Letter to Blok	1914
2	Achmatova, Anna Andreevna	Letter to Brodsky	1965
3	Achmatova, Anna Andreevna	Letter to Chardžiev	1932
4-6	Achmatova, Anna Andreevna	Letter to Chardžiev	1933
7-8	Achmatova, Anna Andreevna	Letter to Chardžiev	1942
9-13	Achmatova, Anna Andreevna	Letter to Chardžiev	1943
14	Achmatova, Anna Andreevna	Letter to Chardžiev	1954
15	Achmatova, Anna Andreevna	Letter to Chardžiev	1965
16	Achmatova, Anna Andreevna	Letter to Čulkov	1914
17	Achmatova, Anna Andreevna	Letter to Čulkov	1930
18-19	Achmatova, Anna Andreevna	Letter to Gumilev	1914
20	Achmatova, Anna Andreevna	Letter to Maksimov	1963

Table A.7: Sources of the Russian texts

Text No.	Author	Title	Year
21	Achmatova, Anna Andreevna	Letter to Mandelštam	1935
22	Achmatova, Anna Andreevna	Letter to Najman	1960
23-25	Achmatova, Anna Andreevna	Letter to Štejn	1906
26-30	Achmatova, Anna Andreevna	Letter to Štejn	1907
31	Journal Novaja Gazeta	Čistyj užas	2002
32	Journal Novaja Gazeta	Ded pogibaet	2002
33	Journal Novaja Gazeta	God kreona	2002
34	Journal Novaja Gazeta	Ich edjat, a oni gljadjat	2002
35	Journal Novaja Gazeta	Igra bez pravil	2002
36	Journal Novaja Gazeta	Istorija ledjanoj	2002
37	Journal Novaja Gazeta	Kurortnaja ljubov	2002
38	Journal Novaja Gazeta	Licedej, kak on est	2002
39	Journal Novaja Gazeta	Malčiki napravo, devočki nalevo	2002
40	Journal Novaja Gazeta	Molodežnyj gus	2002
41	Journal Novaja Gazeta	Nenormativnaja premera	2002
42	Journal Novaja Gazeta	Oblomovskij štolc	2002
43	Journal Novaja Gazeta	Očen smešnoe	2002
44	Journal Novaja Gazeta	Pervaja ledi sssr	2002
45	Journal Novaja Gazeta	Pervaja vysadka	2002
46	Journal Novaja Gazeta	Planeta ljubvi	2002
47	Journal Novaja Gazeta	Povest o poterjannom vremeni	2002
48	Journal Novaja Gazeta	Premera v teatre	2002
49	Journal Novaja Gazeta	Puškin i zlaja baba	2002
50	Journal Novaja Gazeta	Putešestvie Diletantov	2002
51	Journal Novaja Gazeta	Reinkarnacija revizora	2002
52	Journal Novaja Gazeta	Rekonstrukcija ljubvi	2002
53	Journal Novaja Gazeta	Semero svjatych	2002
54	Journal Novaja Gazeta	Skazka o ščaste	2002
55	Journal Novaja Gazeta	Smert todero	2002
56	Journal Novaja Gazeta	Šou kak putešestvie	2002
57	Journal Novaja Gazeta	Sovremmenyj teatr	2002
58	Journal Novaja Gazeta	Spoem tolstogo	2002
59	Journal Novaja Gazeta	Teatr ermitaž v poiskah odessy	2002
60	Journal Novaja Gazeta	Ustalost kultury i golyj zad	2002
61	Nekrasov, Nikolaj Alekseevič	Ach byli ščastlivye gody	1851
62	Nekrasov, Nikolaj Alekseevič	Blazen nezlobivyj poet	1852
63	Nekrasov, Nikolaj Alekseevič	Činovnik	1844
64	Nekrasov, Nikolaj Alekseevič	Da, naša žizn tekla mjatežno	1850
65	Nekrasov, Nikolaj Alekseevič	Edu li nočju po ulice temnoj	1847
66	Nekrasov, Nikolaj Alekseevič	Esli, mušimyj ctrastju mjatežnoj	1847
67	Nekrasov, Nikolaj Alekseevič	Filantrop	1853
68	Nekrasov, Nikolaj Alekseevič	Ja za to gluboko prezuraju sebja	1845
69	Nekrasov, Nikolaj Alekseevič	Kogda iz mraka zabluzdenija	1846
70	Nekrasov, Nikolaj Alekseevič	Kolybelnaja pesnja	1845
71	Nekrasov, Nikolaj Alekseevič	Maša	1855
72	Nekrasov, Nikolaj Alekseevič	Moe razočarovane	1851
73	Nekrasov, Nikolaj Alekseevič	Muza	1852
74	Nekrasov, Nikolaj Alekseevič	Nesžataja polosa	1854
75	Nekrasov, Nikolaj Alekseevič	O pisma ženšiny	1852

Table A.7: Sources of the Russian texts

Text No.	Author	Title	Year
76	Nekrasov, Nikolaj Alekseevič	Ogorodnik	1846
77	Nekrasov, Nikolaj Alekseevič	Otradno videt, čto nachodit	1845
78	Nekrasov, Nikolaj Alekseevič	Otryvok	1844
79	Nekrasov, Nikolaj Alekseevič	Pjanica	1845
80	Nekrasov, Nikolaj Alekseevič	Poražena poterej nevozvgatnoj	1848
81	Nekrasov, Nikolaj Alekseevič	Priznanija truzenika	1854
82	Nekrasov, Nikolaj Alekseevič	Rodina	1846
83	Nekrasov, Nikolaj Alekseevič	Sovremennaja oda	1846
84	Nekrasov, Nikolaj Alekseevič	Stiski! Stiski!	1845
85	Nekrasov, Nikolaj Alekseevič	Svadba	1855
86	Nekrasov, Nikolaj Alekseevič	Tak eto šutka	1850
87	Nekrasov, Nikolaj Alekseevič	Trojka	1846
88	Nekrasov, Nikolaj Alekseevič	V doroge	1845
89	Nekrasov, Nikolaj Alekseevič	V nevedomoj gluši	1846
90	Nekrasov, Nikolaj Alekseevič	Zastenčivost	1852
91	Čechov, Anton Pavlovič	Chameleon	1884
92	Čechov, Anton Pavlovič	Chirurgija	1884
93	Čechov, Anton Pavlovič	Dačnica	1884
94	Čechov, Anton Pavlovič	Dvoe v odnom	1883
95	Čechov, Anton Pavlovič	Edinstvennoe sredstvo	1883
96	Čechov, Anton Pavlovič	Ekzamen na čin	1884
97	Čechov, Anton Pavlovič	Idillija	1884
98	Čechov, Anton Pavlovič	Ispoved	1883
99	Čechov, Anton Pavlovič	Iz ognja da v polymja	1884
100	Čechov, Anton Pavlovič	Na gvozde	1883
101	Čechov, Anton Pavlovič	Na magnetičeskom seanse	1883
102	Čechov, Anton Pavlovič	Nadležasie mery	1884
103	Čechov, Anton Pavlovič	Nevidimye miru slezy	1884
104	Čechov, Anton Pavlovič	Papaša	1880
105	Čechov, Anton Pavlovič	Pered svadboj	1880
106	Čechov, Anton Pavlovič	Petrov den	1881
107	Čechov, Anton Pavlovič	Po amerikanski	1880
108	Čechov, Anton Pavlovič	Radost	1883
109	Čechov, Anton Pavlovič	Rjaženye	1883
110	Čechov, Anton Pavlovič	Russkij ugol	1884
111	Čechov, Anton Pavlovič	Salon de varete	1881
112	Čechov, Anton Pavlovič	Sovet	1883
113	Čechov, Anton Pavlovič	Sovremennye molitvy	1883
114	Čechov, Anton Pavlovič	Sud	1881
115	Čechov, Anton Pavlovič	Temnoju nočju	1883
116	Čechov, Anton Pavlovič	Temperamenty	1881
117	Čechov, Anton Pavlovič	Ušla	1883
118	Čechov, Anton Pavlovič	V cirulne	1883
119	Čechov, Anton Pavlovič	V vagone	1881
120	Čechov, Anton Pavlovič	Za jabločki	1880

Table A.8: Russian texts - frequency distribution and characteristic statistical measures

Text	Frequency class													
No.	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	TL	$\bar{x}$	s^2	d
1	34	27	14	6	2	1	0	0	0	0	84	2.024	1.277	1.247
2	31	54	19	21	4	0	1	0	0	0	130	2.362	1.380	1.013
3	52	44	24	17	7	0	0	0	0	0	144	2.188	1.412	1.189
4	24	19	16	5	0	1	0	0	0	0	65	2.092	1.179	1.079
5	23	26	17	7	0	0	0	0	0	0	73	2.110	0.932	0.840
6	19	14	13	7	1	0	0	0	0	0	54	2.204	1.260	1.046
7	29	14	11	4	0	0	0	0	0	0	58	1.828	0.952	1.151
8	25	13	8	12	3	2	0	0	0	0	63	2.381	2.111	1.528
9	50	36	40	18	3	4	0	0	0	0	151	2.338	1.572	1.175
10	83	82	51	27	9	0	0	0	0	0	252	2.194	1.249	1.046
11	31	28	21	4	3	0	0	0	0	0	87	2.080	1.098	1.016
12	18	17	13	11	4	1	0	0	0	0	64	2.516	1.746	1.152
13	14	11	16	4	1	1	1	0	0	0	48	2.458	1.828	1.253
14	13	27	15	8	1	0	0	1	0	0	65	2.415	1.465	1.035
15	24	16	8	8	0	0	0	0	0	0	56	2.000	1.164	1.164
16	77	54	50	18	2	0	0	0	0	0	201	2.075	1.079	1.004
17	19	16	18	6	4	0	0	0	0	0	63	2.365	1.429	1.047
18	32	52	30	12	3	0	0	0	0	0	129	2.240	1.012	0.816
19	41	57	39	11	2	0	0	0	0	0	150	2.173	0.923	0.786
20	24	21	14	14	2	0	0	0	0	0	75	2.320	1.410	1.068
21	33	26	18	9	2	1	0	0	0	0	89	2.146	1.353	1.181
22	145	154	103	52	25	6	3	0	0	0	488	2.361	1.619	1.190
23	149	96	71	22	4	1	0	0	0	0	343	1.948	1.050	1.108
24	106	100	64	25	6	3	0	0	0	0	304	2.125	1.212	1.077
25	94	89	40	28	7	1	2	0	0	0	261	2.142	1.422	1.246
26	58	42	22	5	2	0	0	0	0	0	129	1.845	0.898	1.062
27	69	63	41	19	4	1	1	0	0	0	198	2.157	1.310	1.133
28	156	139	86	38	14	4	0	0	0	0	437	2.146	1.318	1.150
29	129	113	75	32	16	1	0	0	0	0	366	2.169	1.319	1.128
30	87	106	49	35	14	5	0	0	0	0	296	2.318	1.533	1.163
31	124	132	101	69	37	13	5	1	1	0	483	2.654	2.169	1.311
32	135	147	131	81	31	7	2	1	0	0	535	2.551	1.694	1.092
33	127	166	106	69	28	11	2	2	0	0	511	2.523	1.803	1.184
34	105	114	110	66	33	11	5	0	0	0	444	2.687	1.958	1.161
35	125	135	136	80	29	9	4	1	0	0	519	2.617	1.785	1.104
36	119	123	109	80	42	10	1	1	0	0	485	2.674	1.914	1.143
37	87	122	79	34	21	5	0	0	0	0	348	2.411	1.476	1.046
38	124	127	78	55	21	5	4	0	0	0	414	2.403	1.752	1.249
39	131	160	159	59	27	10	2	1	0	0	549	2.515	1.586	1.047
40	80	86	73	50	23	7	1	0	0	0	320	2.609	1.812	1.126
41	115	112	82	56	33	10	4	0	1	0	413	2.593	2.140	1.343
42	117	139	121	67	20	9	6	0	0	0	479	2.551	1.746	1.125
43	110	137	106	68	34	9	3	0	1	1	469	2.640	1.996	1.217
44	164	176	158	94	55	19	2	1	0	0	669	2.656	1.939	1.170
45	139	128	109	63	27	10	3	0	0	0	479	2.484	1.811	1.220
46	145	144	118	64	34	12	1	1	0	0	519	2.505	1.826	1.213
47	175	178	145	107	47	17	5	0	1	0	675	2.630	2.002	1.229

Table A.8: Russian texts - frequency distribution and characteristic statistical measures

Text	Frequency class													
No.	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	TL	$\bar{x}$	s^2	d
48	120	133	104	51	32	9	5	0	0	0	454	2.535	1.887	1.229
49	172	186	163	78	22	8	3	1	0	0	633	2.420	1.538	1.083
50	110	128	110	67	37	17	0	0	1	0	470	2.681	1.970	1.172
51	95	103	127	86	36	12	4	1	0	0	464	2.832	1.937	1.057
52	230	272	192	127	53	23	3	1	0	0	901	2.542	1.802	1.169
53	208	229	195	123	64	20	2	0	1	0	842	2.620	1.863	1.150
54	82	104	86	30	16	6	0	0	0	0	324	2.420	1.458	1.027
55	149	156	137	82	36	12	5	0	0	0	577	2.577	1.845	1.170
56	90	101	89	51	22	6	0	0	0	0	359	2.532	1.607	1.049
57	208	233	196	109	44	14	4	0	0	0	808	2.512	1.670	1.104
58	190	201	181	114	48	16	5	1	0	0	756	2.606	1.852	1.154
59	128	113	107	53	20	9	1	1	1	0	433	2.460	1.828	1.252
60	170	169	109	78	43	14	7	0	0	0	590	2.534	2.052	1.338
61	25	39	20	9	1	0	0	0	0	0	94	2.170	0.938	0.802
62	43	71	22	16	5	0	0	0	0	0	157	2.166	1.088	0.933
63	395	307	196	98	17	0	1	0	0	0	1014	2.052	1.143	1.087
64	108	84	54	16	6	0	0	0	0	0	268	1.985	1.048	1.064
65	89	96	52	22	3	0	0	0	0	0	262	2.061	0.977	0.921
66	26	39	23	10	3	0	0	0	0	0	101	2.257	1.093	0.869
67	219	151	121	84	34	3	0	0	0	0	612	2.301	1.595	1.226
68	60	27	12	8	2	0	0	0	0	0	109	1.761	1.072	1.408
69	39	46	21	9	4	1	0	0	0	0	120	2.133	1.226	1.082
70	53	65	37	2	2	0	0	0	0	0	159	1.962	0.733	0.761
71	68	67	51	12	3	0	0	0	0	0	201	2.080	0.964	0.893
72	157	139	110	33	3	0	0	0	0	0	442	2.063	0.962	0.905
73	79	81	93	42	6	1	0	0	0	0	302	2.397	1.204	0.861
74	54	63	29	11	1	0	0	0	0	0	158	2.000	0.866	0.866
75	38	45	20	6	2	0	0	0	0	0	111	2.000	0.909	0.909
76	146	146	77	21	3	0	0	0	0	0	393	1.954	0.850	0.891
77	41	42	18	5	1	0	0	0	0	0	107	1.907	0.821	0.906
78	60	52	43	20	7	2	0	0	0	0	184	2.283	1.439	1.122
79	47	44	23	15	6	0	0	0	0	0	135	2.178	1.341	1.139
80	27	24	16	7	3	0	0	0	0	0	77	2.156	1.265	1.094
81	135	157	82	29	5	0	0	0	0	0	408	2.049	0.926	0.883
82	154	143	79	27	7	1	0	0	0	0	411	2.010	1.024	1.014
83	40	29	22	15	2	1	0	0	0	0	109	2.202	1.403	1.168
84	87	70	33	19	3	1	0	0	0	0	213	1.986	1.127	1.143
85	59	68	52	16	1	0	0	0	0	0	196	2.143	0.923	0.808
86	83	72	38	19	5	0	0	0	0	0	217	2.037	1.119	1.079
87	71	80	42	22	2	0	0	0	0	0	217	2.097	1.004	0.916
88	163	145	76	24	3	0	0	0	0	0	411	1.927	0.878	0.947
89	55	54	46	14	3	1	0	0	0	0	173	2.185	1.128	0.952
90	76	76	44	30	8	0	0	0	0	0	234	2.222	1.289	1.055
91	291	275	187	84	29	4	0	0	0	0	870	2.192	1.274	1.069
92	314	292	188	83	22	4	1	0	0	0	904	2.140	1.222	1.071
93	187	211	149	66	39	12	2	0	1	0	667	2.414	1.693	1.198
94	205	188	112	50	14	2	0	1	0	0	572	2.110	1.233	1.111
95	244	217	149	55	28	6	1	0	0	0	700	2.183	1.377	1.164

Table A.8: Russian texts - frequency distribution and characteristic statistical measures

Text	**Frequency class**													
No.	$\mathbf{f_1}$	$\mathbf{f_2}$	$\mathbf{f_3}$	$\mathbf{f_4}$	$\mathbf{f_5}$	$\mathbf{f_6}$	$\mathbf{f_7}$	$\mathbf{f_8}$	$\mathbf{f_9}$	$\mathbf{f_{10}}$	**TL**	$\bar{\mathbf{x}}$	$\mathbf{s^2}$	**d**
96	256	218	215	110	42	16	6	1	1	0	865	2.476	1.833	1.242
97	64	63	57	32	15	5	0	0	0	0	236	2.517	1.706	1.125
98	299	296	182	72	24	3	0	0	0	0	876	2.127	1.164	1.034
99	657	499	347	169	64	10	3	0	0	0	1749	2.157	1.400	1.210
100	115	152	86	38	12	4	1	0	0	0	408	2.255	1.286	1.025
101	201	186	152	53	26	5	0	0	0	0	623	2.249	1.351	1.082
102	261	247	194	121	42	11	3	0	0	0	879	2.410	1.618	1.148
103	452	425	308	132	38	6	0	1	0	0	1362	2.194	1.257	1.053
104	620	424	342	151	53	11	1	1	0	0	1603	2.148	1.401	1.220
105	520	460	266	89	37	14	1	0	0	0	1387	2.069	1.239	1.159
106	880	1021	788	357	72	33	0	2	1	0	3154	2.314	1.295	0.986
107	148	143	118	50	20	5	1	0	1	0	486	2.333	1.538	1.154
108	120	129	127	52	17	2	0	0	0	0	447	2.380	1.290	0.935
109	134	143	110	43	15	4	0	0	0	0	449	2.274	1.298	1.019
110	378	376	231	121	36	11	4	0	0	0	1157	2.231	1.411	1.147
111	269	302	205	84	26	12	2	0	0	0	900	2.267	1.359	1.073
112	146	112	66	34	11	3	0	0	0	0	372	2.089	1.326	1.218
113	97	96	74	44	5	4	0	1	0	0	321	2.318	1.436	1.090
114	334	332	229	97	32	7	1	0	0	0	1032	2.211	1.290	1.065
115	78	94	61	30	9	2	4	0	0	0	278	2.353	1.579	1.168
116	271	240	248	125	55	13	3	0	0	0	955	2.481	1.652	1.116
117	159	110	82	30	6	2	0	0	0	0	389	2.023	1.152	1.125
118	252	265	193	90	23	3	1	0	0	0	827	2.250	1.251	1.000
119	455	448	283	104	53	16	3	1	0	0	1363	2.205	1.425	1.182
120	530	566	365	180	64	12	7	0	0	0	1724	2.273	1.410	1.108

Table A.9: Singh-Poisson model - estimation results for Russian texts

		1-displaced SP				**size-biased SP**
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{MM}} = \hat{\theta}_{\mathrm{ML}}$
1	1.247	0.815	1.256	0.850	1.204	1.024
2	1.013	0.996	1.367	1.046	1.301	1.362
3	1.189	0.868	1.368	0.848	1.401	1.188
4	1.079	0.946	1.155	0.895	1.221	1.092
5	0.840	1.183	0.938	1.050	1.057	1.110
6	1.046	0.978	1.231	0.861	1.398	1.204
7	1.151	0.864	0.958	0.746	1.109	0.828
8	1.528	0.733	1.885	0.701	1.970	1.381
9	1.175	0.889	1.505	0.839	1.594	1.338
10	1.046	0.966	1.236	0.925	1.292	1.194
11	1.016	0.996	1.085	0.945	1.144	1.080
12	1.152	0.919	1.649	0.872	1.738	1.516
13	1.253	0.865	1.686	0.872	1.672	1.458
14	1.035	0.986	1.435	1.110	1.275	1.415
15	1.164	0.875	1.143	0.802	1.247	1.000

Table A.9: Singh-Poisson model - estimation results for Russian texts

		1-displaced SP				**size-biased SP**
No.	d	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\theta}_{\text{MM}} = \hat{\theta}_{\text{ML}}$
16	1.004	1.001	1.074	0.870	1.236	1.075
17	1.047	0.978	1.395	0.891	1.532	1.365
18	0.816	1.181	1.050	1.127	1.101	1.240
19	0.786	1.229	0.955	1.118	1.049	1.173
20	1.068	0.961	1.374	0.872	1.514	1.320
21	1.181	0.872	1.314	0.850	1.349	1.146
22	1.190	0.879	1.548	0.903	1.507	1.361
23	1.108	0.900	1.052	0.832	1.139	0.948
24	1.077	0.938	1.199	0.926	1.215	1.125
25	1.246	0.826	1.383	0.881	1.296	1.142
26	1.062	0.940	0.899	0.910	0.929	0.845
27	1.133	0.901	1.284	0.901	1.283	1.157
28	1.150	0.886	1.293	0.886	1.294	1.146
29	1.128	0.903	1.294	0.881	1.327	1.169
30	1.163	0.892	1.477	0.934	1.411	1.318
31	1.311	0.843	1.962	0.876	1.889	1.654
32	1.092	0.945	1.641	0.916	1.693	1.551
33	1.184	0.893	1.704	0.935	1.628	1.523
34	1.161	0.914	1.845	0.903	1.868	1.687
35	1.104	0.941	1.719	0.916	1.765	1.617
36	1.143	0.922	1.815	0.891	1.880	1.674
37	1.046	0.970	1.454	0.985	1.432	1.411
38	1.249	0.851	1.649	0.878	1.598	1.403
39	1.047	0.971	1.560	0.959	1.581	1.515
40	1.126	0.929	1.732	0.901	1.786	1.609
41	1.343	0.824	1.933	0.854	1.867	1.593
42	1.125	0.926	1.674	0.932	1.664	1.551
43	1.217	0.884	1.854	0.920	1.781	1.640
44	1.170	0.908	1.825	0.896	1.849	1.656
45	1.220	0.872	1.702	0.866	1.715	1.484
46	1.213	0.877	1.716	0.880	1.711	1.505
47	1.229	0.878	1.856	0.878	1.856	1.630
48	1.229	0.871	1.762	0.898	1.709	1.535
49	1.083	0.946	1.502	0.931	1.526	1.420
50	1.172	0.908	1.851	0.909	1.849	1.681
51	1.057	0.971	1.887	0.921	1.988	1.832
52	1.169	0.902	1.709	0.914	1.687	1.542
53	1.150	0.916	1.768	0.903	1.793	1.620
54	1.027	0.984	1.443	0.973	1.459	1.420
55	1.170	0.904	1.745	0.896	1.761	1.577
56	1.049	0.971	1.578	0.927	1.653	1.532
57	1.104	0.936	1.615	0.921	1.643	1.512
58	1.154	0.914	1.758	0.900	1.785	1.606
59	1.252	0.854	1.709	0.864	1.690	1.460
60	1.338	0.820	1.870	0.853	1.798	1.534
61	0.802	1.214	0.964	1.149	1.018	1.170
62	0.933	1.067	1.093	1.126	1.035	1.166
63	1.087	0.925	1.138	0.870	1.209	1.052

Table A.9: Singh-Poisson model - estimation results for Russian texts

		1-displaced SP				**size-biased SP**
No.	d	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\theta}_{\text{MM}} = \hat{\theta}_{\text{ML}}$
64	1.064	0.942	1.045	0.894	1.102	0.985
65	0.921	1.084	0.978	1.023	1.038	1.061
66	0.869	1.125	1.118	1.079	1.165	1.257
67	1.226	0.853	1.525	0.799	1.628	1.301
68	1.408	0.658	1.157	0.653	1.166	0.761
69	1.082	0.940	1.206	0.990	1.144	1.133
70	0.761	1.338	0.719	1.226	0.785	0.962
71	0.893	1.116	0.968	1.005	1.074	1.080
72	0.905	1.101	0.966	0.966	1.100	1.063
73	0.861	1.113	1.256	0.966	1.447	1.397
74	0.866	1.162	0.861	1.106	0.904	1.000
75	0.909	1.110	0.901	1.104	0.906	1.000
76	0.891	1.132	0.843	1.057	0.902	0.954
77	0.906	1.127	0.804	1.097	0.827	0.907
78	1.122	0.917	1.398	0.877	1.462	1.283
79	1.139	0.900	1.308	0.887	1.328	1.178
80	1.094	0.935	1.236	0.896	1.290	1.156
81	0.883	1.128	0.930	1.072	0.978	1.049
82	1.014	0.988	1.022	0.962	1.049	1.010
83	1.168	0.884	1.359	0.826	1.456	1.202
84	1.143	0.877	1.124	0.875	1.126	0.986
85	0.808	1.208	0.946	1.059	1.080	1.143
86	1.079	0.933	1.111	0.906	1.145	1.037
87	0.916	1.088	1.008	1.023	1.072	1.097
88	0.947	1.064	0.871	0.996	0.930	0.927
89	0.952	1.047	1.132	0.964	1.229	1.185
90	1.055	0.960	1.273	0.917	1.333	1.222
91	1.069	0.946	1.259	0.913	1.306	1.192
92	1.071	0.942	1.210	0.917	1.244	1.140
93	1.198	0.878	1.610	0.915	1.546	1.414
94	1.111	0.911	1.219	0.911	1.219	1.110
95	1.164	0.879	1.345	0.882	1.341	1.183
96	1.242	0.860	1.717	0.857	1.722	1.476
97	1.125	0.927	1.637	0.891	1.702	1.517
98	1.034	0.972	1.159	0.947	1.190	1.127
99	1.210	0.847	1.367	0.831	1.393	1.157
100	1.025	0.982	1.277	1.009	1.244	1.255
101	1.082	0.940	1.329	0.905	1.380	1.249
102	1.148	0.906	1.556	0.881	1.600	1.410
103	1.053	0.958	1.246	0.919	1.300	1.194
104	1.220	0.840	1.368	0.809	1.420	1.148
105	1.159	0.871	1.227	0.898	1.190	1.069
106	0.986	1.011	1.299	0.974	1.349	1.314
107	1.154	0.898	1.485	0.900	1.481	1.333
108	0.935	1.051	1.313	0.959	1.440	1.380
109	1.019	0.987	1.290	0.950	1.341	1.274
110	1.147	0.894	1.376	0.906	1.358	1.231
111	1.073	0.946	1.339	0.954	1.328	1.267

Table A.9: Singh-Poisson model - estimation results for Russian texts

		1-displaced SP				**size-biased SP**
No.	d	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\theta}_{\text{MM}} = \hat{\theta}_{\text{ML}}$
112	1.218	0.835	1.304	0.833	1.307	1.089
113	1.090	0.938	1.404	0.914	1.442	1.318
114	1.065	0.950	1.275	0.928	1.305	1.211
115	1.168	0.892	1.516	0.946	1.430	1.353
116	1.116	0.928	1.595	0.880	1.683	1.481
117	1.125	0.893	1.146	0.839	1.219	1.023
118	1.000	1.001	1.250	0.950	1.316	1.250
119	1.182	0.869	1.386	0.905	1.332	1.205
120	1.108	0.922	1.380	0.928	1.371	1.273

Table A.10: Hyper-Poisson model - estimation results for Russian texts

		1-displaced HP						**size-biased HP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
1	1.247	2.203	1.750	2.344	1.822	2.398	1.856	5.053	2.475	5.937	2.776	6.356	2.921
2	1.013	1.124	1.436	0.904	1.289	1.042	1.394	1.193	1.442	0.877	1.306	∅	∅
3	1.189	0.967	1.268	2.216	1.974	2.069	1.871	1.781	1.546	4.461	2.558	6.477	3.298
4	1.079	1.057	1.155	1.425	1.365	1.380	1.332	1.326	1.230	1.979	1.468	1.327	1.236
5	0.840	0.011	0.499	0.703	0.903	0.471	0.747	∅	0.440	0.207	0.787	∅	∅
6	1.046	0.139	0.748	1.556	1.577	1.249	1.365	∅	0.766	2.094	1.651	1.180	1.289
7	1.151	0.458	0.641	2.352	1.534	2.446	1.550	0.739	0.789	4.705	1.964	1.260	0.948
8	1.528	1.824	2.151	6.137	4.702	14.224	9.357	9.506	5.061	27.265	12.327	∅	∅
9	1.175	1.130	1.508	2.080	2.084	2.036	2.031	1.884	1.759	3.744	2.519	8.277	4.324
10	1.046	0.599	0.999	1.365	1.442	1.196	1.326	0.535	1.036	1.827	1.533	1.249	1.305
11	1.016	0.711	0.933	1.159	1.184	1.068	1.124	0.586	0.937	1.305	1.198	1.107	1.124
12	1.152	0.567	1.335	1.966	2.227	1.787	2.081	0.780	1.479	3.258	2.572	5.613	3.570
13	1.253	2.511	2.489	1.992	2.175	2.306	2.384	4.555	3.053	3.862	2.758	6.310	3.804
14	1.035	2.789	2.475	0.627	1.121	1.097	1.493	3.622	2.510	0.257	1.061	∅	∅
15	1.164	0.476	0.810	2.254	1.737	2.145	1.655	0.771	0.971	4.380	2.191	1.354	1.173
16	1.004	0.166	0.642	1.400	1.330	1.023	1.089	∅	0.642	1.717	1.348	1.020	1.083
17	1.047	0.347	1.006	1.464	1.699	1.223	1.521	0.141	1.029	1.905	1.768	1.211	1.470
18	0.816	0.277	0.735	0.541	0.895	0.486	0.854	∅	0.648	∅	0.766	∅	∅
19	0.786	0.124	0.582	0.543	0.839	0.383	0.725	∅	0.503	∅	0.705	∅	∅
20	1.068	0.128	0.855	1.671	1.789	1.348	1.557	∅	0.893	2.399	1.927	1.267	1.453
21	1.181	1.266	1.378	2.046	1.814	2.015	1.785	2.250	1.662	3.955	2.293	6.015	3.039
22	1.190	1.690	1.866	1.886	1.983	1.866	1.969	2.939	2.221	3.589	2.489	3.782	2.564
23	1.108	0.868	0.934	1.819	1.422	1.665	1.323	1.267	1.067	3.118	1.673	1.292	1.076
24	1.077	1.121	1.228	1.384	1.377	1.336	1.344	1.471	1.321	1.990	1.513	1.402	1.298
25	1.246	2.434	2.034	2.214	1.911	2.226	1.926	5.257	2.776	5.323	2.800	4.810	2.615
26	1.062	0.982	0.862	1.360	1.045	1.309	1.015	1.215	0.925	1.936	1.143	1.280	0.947
27	1.133	1.521	1.504	1.620	1.563	1.619	1.560	2.401	1.716	2.721	1.837	3.196	2.010
28	1.150	1.338	1.405	1.798	1.662	1.734	1.618	2.229	1.648	3.283	2.038	3.797	2.220
29	1.128	1.001	1.237	1.763	1.669	1.636	1.581	1.514	1.405	3.031	1.975	1.443	1.380
30	1.163	1.454	1.666	1.696	1.804	1.675	1.794	2.388	1.932	3.092	2.215	2.575	2.008

Table A.10: Hyper-Poisson model - estimation results for Russian texts

		1-displaced HP						**size-biased HP**					
No.	d	$\hat{\lambda}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\lambda}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\lambda}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\lambda}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\lambda}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\lambda}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$
31	1.311	2.280	2.644	2.636	2.876	2.614	2.854	4.612	3.429	5.897	4.018	7.473	4.725
32	1.092	0.939	1.569	1.547	1.966	1.399	1.850	1.185	1.671	2.235	2.145	1.448	1.787
33	1.184	1.856	2.149	1.719	2.058	1.752	2.087	3.047	2.486	3.080	2.500	2.789	2.375
34	1.161	1.159	1.886	1.856	2.350	1.744	2.255	1.752	2.103	3.087	2.728	1.673	2.068
35	1.104	1.102	1.743	1.567	2.054	1.456	1.963	1.462	1.869	2.293	2.253	1.505	1.889
36	1.143	0.797	1.636	1.907	2.373	1.698	2.201	1.138	1.801	3.101	2.718	1.572	2.000
37	1.046	1.001	1.437	1.152	1.524	1.164	1.534	1.132	1.483	1.423	1.605	∅	∅
38	1.249	1.896	2.070	2.296	2.312	2.245	2.275	3.722	2.627	4.945	3.138	5.609	3.406
39	1.047	1.163	1.641	1.204	1.672	1.176	1.650	1.365	1.692	1.453	1.733	1.319	1.673
40	1.126	0.787	1.553	1.773	2.198	1.584	2.047	1.065	1.693	2.803	2.488	1.541	1.907
41	1.343	2.334	2.622	2.984	3.036	2.967	3.013	5.180	3.597	7.259	4.528	10.808	6.094
42	1.125	1.459	1.905	1.522	1.948	1.517	1.942	2.099	2.085	2.332	2.190	2.787	2.387
43	1.217	2.351	2.612	1.868	2.304	1.935	2.356	3.973	3.075	3.430	2.829	3.657	2.934
44	1.170	1.086	1.815	1.951	2.384	1.809	2.267	1.694	2.047	3.351	2.814	4.684	3.402
45	1.220	1.434	1.873	2.242	2.380	2.170	2.315	2.607	2.262	4.392	3.034	6.698	4.005
46	1.213	1.484	1.922	2.146	2.339	2.069	2.275	2.634	2.297	4.132	2.951	5.551	3.550
47	1.229	1.513	2.088	2.255	2.573	2.170	2.497	2.691	2.491	4.327	3.239	6.343	4.131
48	1.229	1.792	2.152	2.097	2.343	2.067	2.320	3.231	2.602	4.157	3.011	4.592	3.195
49	1.083	1.251	1.615	1.397	1.712	1.345	1.671	1.640	1.720	1.941	1.850	1.454	1.643
50	1.172	1.299	1.975	1.873	2.356	1.777	2.276	2.021	2.225	3.193	2.772	4.068	3.160
51	1.057	0.677	1.658	1.481	2.223	1.250	2.031	0.693	1.711	1.931	2.327	1.302	2.004
52	1.169	1.354	1.858	1.820	2.154	1.741	2.094	2.152	2.116	3.179	2.572	3.551	2.726
53	1.150	1.099	1.773	1.826	2.250	1.688	2.138	1.638	1.972	3.015	2.603	1.630	1.969
54	1.027	0.929	1.392	1.132	1.519	1.099	1.493	0.967	1.416	1.299	1.558	1.204	1.516
55	1.170	1.255	1.836	1.900	2.253	1.801	2.172	1.983	2.084	3.282	2.670	4.458	3.177
56	1.049	0.566	1.298	1.412	1.845	1.206	1.686	0.509	1.339	1.866	1.947	1.267	1.669
57	1.104	1.033	1.594	1.558	1.931	1.447	1.844	1.379	1.720	2.340	2.145	1.518	1.780
58	1.154	1.179	1.810	1.830	2.237	1.714	2.141	1.780	2.021	3.036	2.594	1.627	1.953
59	1.252	2.273	2.363	2.271	2.368	2.349	2.410	4.345	2.976	4.715	3.136	7.091	4.136
60	1.338	2.123	2.417	3.019	2.977	2.921	2.901	4.866	3.358	7.510	4.516	10.187	5.668
61	0.802	0.209	0.632	0.514	0.813	0.452	0.768	∅	0.550	∅	0.686	∅	∅
62	0.933	0.794	1.015	0.654	0.918	0.809	1.027	0.512	0.959	0.257	0.854	0.437	0.934
63	1.087	0.750	0.968	1.653	1.458	1.454	1.330	0.942	1.062	2.611	1.650	1.305	1.187
64	1.064	0.804	0.918	1.456	1.261	1.317	1.174	0.930	0.982	2.116	1.383	2.721	1.574
65	0.921	0.382	0.700	0.849	0.961	0.720	0.876	∅	0.653	0.565	0.892	0.413	0.830
66	0.869	0.348	0.817	0.662	1.006	0.608	0.966	∅	0.743	0.187	0.897	0.138	0.874
67	1.226	0.791	1.310	2.709	2.440	2.667	2.371	1.674	1.655	5.725	3.271	27.766	11.937
68	1.408	3.534	2.028	6.947	3.518	11.227	5.359	∅	∅	∅	∅	∅	∅
69	1.082	1.400	1.384	1.178	1.252	1.301	1.337	1.903	1.486	1.615	1.377	∅	∅
70	0.761	0.248	0.474	0.438	0.585	0.336	0.520	∅	0.395	∅	0.471	∅	∅
71	0.893	0.245	0.630	0.836	0.970	0.609	0.821	∅	0.576	0.487	0.877	0.255	0.781
72	0.905	0.114	0.556	0.944	1.027	0.620	0.819	∅	0.513	0.693	0.944	0.350	0.805
73	0.861	∅	0.689	0.910	1.329	0.481	1.014	∅	0.631	0.587	1.204	0.147	0.990
74	0.866	0.350	0.605	0.644	0.765	0.577	0.722	∅	0.539	0.130	0.670	0.061	0.642
75	0.909	0.639	0.770	0.676	0.788	0.712	0.810	0.217	0.706	0.227	0.708	0.288	0.732
76	0.891	0.369	0.591	0.746	0.793	0.620	0.715	∅	0.535	0.327	0.710	0.190	0.657
77	0.906	0.545	0.642	0.695	0.718	0.684	0.712	0.064	0.583	0.244	0.641	0.251	0.644
78	1.122	0.894	1.288	1.750	1.798	1.611	1.694	1.261	1.430	2.851	2.065	1.434	1.500

Table A.10: Hyper-Poisson model - estimation results for Russian texts

		1-displaced HP						**size-biased HP**					
No.	d	$\hat\lambda_{\mathrm{MM}}$	$\hat\theta_{\mathrm{MM}}$	$\hat\lambda_{\mathrm{ML}}$	$\hat\theta_{\mathrm{ML}}$	$\hat\lambda_{\mathrm{FF}}$	$\hat\theta_{\mathrm{FF}}$	$\hat\lambda_{\mathrm{MM}}$	$\hat\theta_{\mathrm{MM}}$	$\hat\lambda_{\mathrm{ML}}$	$\hat\theta_{\mathrm{ML}}$	$\hat\lambda_{\mathrm{FF}}$	$\hat\theta_{\mathrm{FF}}$
79	1.139	0.914	1.202	1.806	1.706	1.676	1.619	1.395	1.372	3.179	2.045	1.483	1.408
80	1.094	0.747	1.063	1.559	1.524	1.447	1.446	0.900	1.148	2.366	1.698	1.383	1.331
81	0.883	0.389	0.675	0.711	0.855	0.619	0.794	∅	0.613	0.270	0.764	0.169	0.723
82	1.014	0.827	0.929	1.130	1.091	1.059	1.047	0.780	0.940	1.285	1.115	1.110	1.051
83	1.168	0.799	1.178	2.162	1.960	2.050	1.866	1.333	1.386	4.014	2.409	1.427	1.425
84	1.143	1.294	1.200	1.814	1.470	1.748	1.428	2.217	1.433	3.430	1.840	3.864	1.979
85	0.808	∅	0.500	0.658	0.900	0.381	0.710	∅	0.436	0.109	0.771	∅	∅
86	1.079	0.828	0.987	1.491	1.342	1.372	1.266	1.022	1.070	2.263	1.503	1.346	1.182
87	0.916	0.315	0.692	0.848	0.994	0.706	0.899	∅	0.644	0.561	0.922	0.389	0.850
88	0.947	0.460	0.641	0.934	0.887	0.790	0.801	0.043	0.606	0.749	0.839	0.579	0.777
89	0.952	0.465	0.866	1.010	1.192	0.815	1.059	0.099	0.830	0.885	1.137	0.678	1.047
90	1.055	0.536	0.995	1.436	1.519	1.242	1.385	0.468	1.040	1.997	1.635	1.281	1.350
91	1.069	0.811	1.126	1.453	1.496	1.307	1.396	0.947	1.197	2.084	1.633	2.621	1.823
92	1.071	0.968	1.161	1.422	1.419	1.316	1.347	1.210	1.242	2.043	1.553	2.481	1.704
93	1.198	1.968	2.096	1.836	2.014	1.865	2.036	3.421	2.498	3.490	2.527	3.464	2.516
94	1.111	1.425	1.391	1.531	1.452	1.504	1.434	2.152	1.560	2.472	1.678	2.766	1.780
95	1.164	1.490	1.536	1.857	1.745	1.812	1.712	2.572	1.827	3.489	2.174	4.116	2.402
96	1.242	1.928	2.173	2.304	2.412	2.336	2.417	3.635	2.697	4.681	3.150	7.840	4.484
97	1.125	0.773	1.452	1.772	2.091	1.596	1.951	1.046	1.589	2.827	2.376	1.500	1.786
98	1.034	0.822	1.049	1.235	1.282	1.138	1.218	0.842	1.081	1.546	1.343	1.220	1.218
99	1.210	1.440	1.506	2.352	2.016	2.282	1.958	2.906	1.932	5.082	2.739	8.555	4.003
100	1.025	1.205	1.381	1.009	1.262	1.086	1.317	1.372	1.407	1.079	1.289	∅	∅
101	1.082	0.902	1.239	1.504	1.595	1.373	1.501	1.131	1.330	2.211	1.756	1.369	1.422
102	1.148	1.034	1.515	1.881	2.039	1.732	1.924	1.589	1.714	3.222	2.404	1.527	1.688
103	1.053	0.852	1.142	1.374	1.447	1.232	1.349	0.955	1.197	1.854	1.543	1.274	1.316
104	1.220	1.448	1.505	2.500	2.092	2.495	2.065	3.033	1.970	5.534	2.893	13.570	5.814
105	1.159	1.787	1.558	1.737	1.530	1.758	1.543	3.182	1.884	3.324	1.935	3.362	1.947
106	0.986	0.802	1.193	1.052	1.352	0.947	1.276	0.677	1.180	1.047	1.335	0.900	1.268
107	1.154	1.774	1.862	1.697	1.823	1.708	1.825	2.854	2.135	2.879	2.148	3.625	2.445
108	0.935	0.250	0.906	1.032	1.405	0.750	1.197	∅	0.866	0.909	1.338	0.586	1.185
109	1.019	0.745	1.140	1.194	1.412	1.073	1.325	0.678	1.155	1.396	1.444	1.121	1.328
110	1.147	1.500	1.587	1.697	1.701	1.665	1.679	2.435	1.833	2.983	2.046	3.219	2.132
111	1.073	1.276	1.461	1.283	1.465	1.287	1.468	1.671	1.555	1.750	1.586	1.899	1.643
112	1.218	1.575	1.500	2.383	1.938	2.318	1.890	3.342	1.990	5.448	2.740	7.864	3.585
113	1.090	1.239	1.503	1.469	1.649	1.395	1.594	1.665	1.616	2.115	1.802	1.426	1.523
114	1.065	0.963	1.226	1.375	1.467	1.277	1.398	1.166	1.298	1.910	1.586	1.353	1.369
115	1.168	2.177	2.123	1.526	1.728	1.667	1.832	3.586	2.462	2.644	2.074	∅	∅
116	1.116	0.802	1.431	1.776	2.050	1.580	1.896	1.078	1.562	2.786	2.308	1.442	1.718
117	1.125	0.985	1.078	1.873	1.551	1.742	1.462	1.541	1.248	3.312	1.858	1.343	1.182
118	1.000	0.611	1.034	1.167	1.368	1.002	1.251	0.436	1.033	1.300	1.378	1.003	1.252
119	1.182	1.949	1.823	1.798	1.735	1.831	1.759	3.503	2.215	3.499	2.213	3.445	2.193
120	1.108	1.309	1.506	1.492	1.614	1.457	1.589	1.886	1.659	2.320	1.832	2.515	1.904

Table A.11: Generalized Poisson model - estimation results for Russian texts

		1-displaced GP						**size-biased GP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
1	1.247	0.104	0.917	0.104	0.917	0.117	0.904	0.099	0.713	0.099	0.712	0.110	0.679
2	1.013	0.007	1.352	0.003	1.358	-0.053	1.434	0.007	1.339	0.003	1.352	-0.054	1.540
3	1.189	0.083	1.089	0.094	1.076	0.142	1.019	0.080	0.925	0.091	0.890	0.133	0.742
4	1.079	0.037	1.051	0.032	1.058	0.088	0.996	0.037	0.977	0.031	0.994	0.084	0.824
5	0.840	-0.091	1.211	-0.116	1.239	-0.041	1.155	-0.095	1.396	-0.121	1.473	-0.042	1.237
6	1.046	0.022	1.177	0.017	1.184	0.132	1.045	0.022	1.132	0.017	1.150	0.125	0.787
7	1.151	0.068	0.772	0.078	0.763	0.162	0.693	0.065	0.639	0.076	0.608	0.146	0.388
8	1.528	0.191	1.117	0.238	1.052	0.331	0.924	0.178	0.742	0.222	0.567	0.285	0.304
9	1.175	0.077	1.234	0.085	1.224	0.174	1.105	0.075	1.081	0.082	1.055	0.161	0.768
10	1.046	0.022	1.168	0.023	1.167	0.070	1.111	0.022	1.124	0.023	1.121	0.068	0.972
11	1.016	0.008	1.072	0.002	1.078	0.045	1.032	0.008	1.056	0.002	1.073	0.044	0.943
12	1.152	0.068	1.412	0.073	1.405	0.163	1.269	0.067	1.276	0.071	1.259	0.153	0.949
13	1.253	0.107	1.303	0.098	1.316	0.155	1.232	0.103	1.091	0.094	1.123	0.146	0.928
14	1.035	0.017	1.391	0.007	1.405	-0.137	1.609	0.017	1.357	0.007	1.391	-0.143	1.886
15	1.164	0.073	0.927	0.082	0.918	0.153	0.847	0.070	0.784	0.080	0.754	0.141	0.555
16	1.004	0.002	1.072	0.000	1.075	0.107	0.959	0.002	1.068	0.000	1.076	0.102	0.751
17	1.047	0.023	1.334	0.018	1.341	0.122	1.199	0.022	1.289	0.017	1.306	0.116	0.959
18	0.816	-0.107	1.373	-0.114	1.381	-0.124	1.394	-0.111	1.590	-0.118	1.611	-0.130	1.646
19	0.786	-0.128	1.323	-0.138	1.336	-0.105	1.297	-0.134	1.584	-0.146	1.617	-0.110	1.511
20	1.068	0.032	1.277	0.033	1.277	0.137	1.139	0.032	1.213	0.032	1.211	0.129	0.872
21	1.181	0.080	1.055	0.084	1.050	0.134	0.992	0.077	0.897	0.081	0.884	0.126	0.731
22	1.190	0.083	1.247	0.087	1.242	0.108	1.214	0.081	1.082	0.084	1.069	0.104	1.001
23	1.108	0.050	0.900	0.056	0.894	0.120	0.834	0.049	0.802	0.055	0.782	0.112	0.602
24	1.077	0.037	1.084	0.037	1.083	0.063	1.054	0.036	1.011	0.036	1.010	0.062	0.928
25	1.246	0.104	1.023	0.105	1.022	0.106	1.021	0.099	0.819	0.100	0.815	0.100	0.815
26	1.062	0.030	0.820	0.028	0.821	0.054	0.799	0.029	0.761	0.028	0.766	0.052	0.694
27	1.133	0.061	1.087	0.061	1.087	0.089	1.054	0.059	0.967	0.059	0.966	0.085	0.880
28	1.150	0.067	1.069	0.072	1.064	0.102	1.030	0.065	0.936	0.070	0.921	0.097	0.831
29	1.128	0.058	1.101	0.064	1.094	0.108	1.043	0.057	0.985	0.063	0.966	0.103	0.831
30	1.163	0.073	1.222	0.076	1.218	0.071	1.224	0.071	1.077	0.074	1.067	0.069	1.085
31	1.311	0.127	1.445	0.136	1.429	0.178	1.360	0.122	1.193	0.131	1.157	0.167	1.009
32	1.092	0.043	1.485	0.046	1.480	0.112	1.377	0.042	1.399	0.045	1.388	0.108	1.155
33	1.184	0.081	1.399	0.082	1.397	0.086	1.392	0.079	1.238	0.080	1.233	0.083	1.222
34	1.161	0.072	1.566	0.078	1.555	0.145	1.442	0.070	1.423	0.076	1.399	0.138	1.155
35	1.104	0.048	1.538	0.051	1.534	0.119	1.424	0.048	1.442	0.050	1.432	0.115	1.187
36	1.143	0.065	1.566	0.074	1.551	0.161	1.405	0.064	1.437	0.072	1.403	0.152	1.088
37	1.046	0.022	1.379	0.022	1.380	0.017	1.386	0.022	1.335	0.022	1.337	0.017	1.351
38	1.249	0.105	1.256	0.112	1.246	0.141	1.206	0.101	1.048	0.108	1.022	0.133	0.929
39	1.047	0.022	1.481	0.022	1.482	0.054	1.433	0.022	1.437	0.022	1.439	0.053	1.324
40	1.126	0.058	1.517	0.064	1.506	0.139	1.386	0.057	1.402	0.063	1.378	0.132	1.113
41	1.343	0.137	1.375	0.150	1.354	0.198	1.279	0.131	1.103	0.144	1.054	0.184	0.892
42	1.125	0.057	1.462	0.058	1.461	0.091	1.410	0.056	1.348	0.057	1.344	0.088	1.228
43	1.217	0.094	1.486	0.094	1.486	0.116	1.450	0.091	1.300	0.091	1.300	0.111	1.221
44	1.170	0.076	1.531	0.084	1.517	0.151	1.406	0.074	1.380	0.082	1.348	0.143	1.108
45	1.220	0.095	1.344	0.105	1.329	0.166	1.237	0.092	1.156	0.101	1.119	0.156	0.911
46	1.213	0.092	1.366	0.101	1.353	0.153	1.275	0.089	1.183	0.098	1.152	0.144	0.975
47	1.229	0.098	1.470	0.108	1.454	0.172	1.350	0.095	1.276	0.105	1.237	0.162	1.012

Table A.11: Generalized Poisson model - estimation results for Russian texts

		1-displaced GP						**size-biased GP**					
No.	d	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\lambda}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\lambda}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\lambda}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$
48	1.229	0.098	1.385	0.104	1.376	0.133	1.331	0.095	1.190	0.101	1.169	0.127	1.068
49	1.083	0.039	1.365	0.040	1.364	0.083	1.303	0.039	1.287	0.039	1.285	0.080	1.140
50	1.172	0.076	1.552	0.082	1.542	0.136	1.452	0.075	1.400	0.081	1.377	0.130	1.183
51	1.057	0.028	1.781	0.029	1.778	0.134	1.586	0.027	1.726	0.029	1.720	0.129	1.319
52	1.169	0.075	1.426	0.081	1.417	0.114	1.365	0.073	1.277	0.079	1.256	0.110	1.139
53	1.150	0.067	1.511	0.074	1.500	0.137	1.398	0.066	1.377	0.073	1.351	0.131	1.128
54	1.027	0.013	1.401	0.012	1.403	0.032	1.374	0.013	1.375	0.012	1.378	0.032	1.310
55	1.170	0.075	1.458	0.082	1.447	0.142	1.354	0.074	1.308	0.080	1.283	0.135	1.075
56	1.049	0.024	1.496	0.025	1.493	0.097	1.384	0.023	1.449	0.025	1.443	0.094	1.191
57	1.104	0.048	1.439	0.052	1.434	0.103	1.357	0.048	1.343	0.051	1.330	0.099	1.154
58	1.154	0.069	1.495	0.075	1.485	0.140	1.381	0.067	1.358	0.074	1.335	0.133	1.105
59	1.252	0.106	1.304	0.111	1.298	0.165	1.219	0.102	1.094	0.107	1.077	0.155	0.896
60	1.338	0.135	1.326	0.150	1.303	0.189	1.244	0.129	1.058	0.143	1.003	0.176	0.875
61	0.802	-0.117	1.307	-0.127	1.319	-0.132	1.324	-0.122	1.545	-0.134	1.578	-0.139	1.593
62	0.933	-0.035	1.207	-0.038	1.210	-0.111	1.295	-0.036	1.277	-0.038	1.285	-0.116	1.521
63	1.087	0.041	1.009	0.046	1.003	0.104	0.943	0.040	0.929	0.046	0.911	0.099	0.740
64	1.064	0.031	0.955	0.032	0.953	0.077	0.909	0.030	0.894	0.032	0.889	0.074	0.757
65	0.921	-0.042	1.106	-0.048	1.112	-0.018	1.080	-0.043	1.191	-0.049	1.210	-0.018	1.115
66	0.869	-0.073	1.349	-0.081	1.360	-0.079	1.357	-0.075	1.495	-0.084	1.524	-0.082	1.517
67	1.226	0.097	1.174	0.119	1.145	0.210	1.028	0.093	0.983	0.115	0.906	0.191	0.624
68	1.408	0.157	0.642	0.183	0.622	0.216	0.597	0.142	0.345	0.165	0.273	0.186	0.205
69	1.082	0.038	1.090	0.034	1.095	0.008	1.124	0.038	1.013	0.034	1.027	0.008	1.107
70	0.761	-0.146	1.103	-0.140	1.097	-0.142	1.099	-0.157	1.406	-0.151	1.390	-0.152	1.391
71	0.893	-0.058	1.143	-0.068	1.153	-0.004	1.084	-0.060	1.261	-0.070	1.290	-0.004	1.092
72	0.905	-0.051	1.118	-0.062	1.130	0.027	1.035	-0.053	1.222	-0.064	1.256	0.026	0.982
73	0.861	-0.077	1.506	-0.094	1.528	0.040	1.341	-0.080	1.662	-0.096	1.715	0.040	1.261
74	0.866	-0.074	1.074	-0.082	1.082	-0.074	1.074	-0.077	1.226	-0.085	1.248	-0.076	1.223
75	0.909	-0.049	1.049	-0.053	1.053	-0.072	1.072	-0.050	1.148	-0.054	1.160	-0.074	1.218
76	0.891	-0.060	1.011	-0.065	1.016	-0.038	0.990	-0.061	1.132	-0.067	1.149	-0.038	1.066
77	0.906	-0.051	0.952	-0.056	0.958	-0.058	0.959	-0.052	1.055	-0.058	1.072	-0.060	1.077
78	1.122	0.056	1.211	0.060	1.205	0.126	1.121	0.055	1.100	0.059	1.085	0.120	0.873
79	1.139	0.063	1.104	0.068	1.098	0.104	1.055	0.061	0.979	0.066	0.963	0.099	0.851
80	1.094	0.044	1.105	0.043	1.106	0.093	1.048	0.043	1.017	0.042	1.020	0.089	0.865
81	0.883	-0.064	1.116	-0.069	1.121	-0.054	1.106	-0.066	1.246	-0.071	1.261	-0.056	1.216
82	1.014	0.007	1.003	0.006	1.003	0.028	0.982	0.007	0.988	0.006	0.991	0.027	0.926
83	1.168	0.075	1.112	0.084	1.101	0.166	1.002	0.072	0.965	0.082	0.932	0.154	0.682
84	1.143	0.065	0.922	0.069	0.918	0.092	0.895	0.063	0.795	0.066	0.783	0.087	0.716
85	0.808	-0.113	1.272	-0.133	1.295	-0.051	1.201	-0.118	1.502	-0.139	1.564	-0.052	1.302
86	1.079	0.037	0.998	0.040	0.996	0.073	0.961	0.037	0.924	0.039	0.917	0.070	0.818
87	0.916	-0.045	1.146	-0.053	1.155	-0.019	1.117	-0.046	1.237	-0.054	1.261	-0.019	1.155
88	0.947	-0.028	0.953	-0.032	0.957	0.002	0.925	-0.028	1.009	-0.032	1.021	0.002	0.920
89	0.952	-0.025	1.214	-0.030	1.221	0.033	1.146	-0.025	1.264	-0.030	1.281	0.032	1.081
90	1.055	0.026	1.190	0.028	1.188	0.080	1.125	0.026	1.137	0.028	1.131	0.077	0.967
91	1.069	0.033	1.153	0.036	1.149	0.081	1.095	0.032	1.088	0.035	1.078	0.078	0.935
92	1.071	0.034	1.102	0.036	1.099	0.073	1.057	0.033	1.035	0.036	1.028	0.070	0.914
93	1.198	0.086	1.292	0.088	1.289	0.101	1.272	0.084	1.121	0.086	1.114	0.097	1.073
94	1.111	0.051	1.053	0.052	1.052	0.076	1.026	0.050	0.952	0.051	0.948	0.073	0.877
95	1.164	0.073	1.096	0.078	1.091	0.109	1.054	0.071	0.952	0.076	0.936	0.104	0.841

Table A.11: Generalized Poisson model - estimation results for Russian texts

		1-displaced GP						**size-biased GP**					
No.	d	$\hat{\lambda}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\lambda}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\lambda}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\lambda}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\lambda}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\lambda}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$
96	1.242	0.103	1.325	0.110	1.313	0.175	1.218	0.099	1.122	0.107	1.093	0.164	0.875
97	1.125	0.057	1.430	0.063	1.421	0.140	1.305	0.056	1.317	0.062	1.295	0.133	1.030
98	1.034	0.016	1.108	0.017	1.107	0.046	1.075	0.016	1.076	0.017	1.074	0.045	0.984
99	1.210	0.091	1.052	0.103	1.038	0.154	0.979	0.087	0.873	0.099	0.835	0.143	0.682
100	1.025	0.012	1.240	0.011	1.241	-0.009	1.266	0.012	1.215	0.011	1.220	-0.009	1.285
101	1.082	0.039	1.201	0.042	1.197	0.094	1.131	0.038	1.124	0.041	1.113	0.090	0.946
102	1.148	0.067	1.316	0.075	1.304	0.139	1.214	0.065	1.183	0.073	1.156	0.131	0.942
103	1.053	0.025	1.163	0.027	1.161	0.076	1.103	0.025	1.113	0.027	1.107	0.074	0.953
104	1.220	0.095	1.040	0.108	1.024	0.173	0.950	0.091	0.854	0.104	0.809	0.159	0.618
105	1.159	0.071	0.993	0.073	0.991	0.082	0.981	0.069	0.853	0.070	0.848	0.079	0.819
106	0.986	-0.007	1.323	-0.008	1.323	0.028	1.277	-0.007	1.337	-0.008	1.339	0.028	1.220
107	1.154	0.069	1.241	0.070	1.240	0.108	1.189	0.067	1.105	0.068	1.101	0.104	0.976
108	0.935	-0.034	1.428	-0.040	1.436	0.047	1.315	-0.035	1.497	-0.041	1.516	0.046	1.221
109	1.019	0.009	1.262	0.009	1.263	0.051	1.209	0.009	1.244	0.009	1.246	0.050	1.108
110	1.147	0.066	1.149	0.069	1.145	0.091	1.119	0.064	1.019	0.068	1.008	0.088	0.940
111	1.073	0.035	1.223	0.035	1.223	0.047	1.208	0.034	1.154	0.034	1.153	0.046	1.115
112	1.218	0.094	0.986	0.104	0.976	0.141	0.935	0.090	0.802	0.100	0.770	0.131	0.663
113	1.090	0.042	1.262	0.043	1.262	0.092	1.197	0.041	1.178	0.042	1.177	0.088	1.016
114	1.065	0.031	1.174	0.033	1.172	0.069	1.128	0.031	1.112	0.032	1.106	0.067	0.993
115	1.168	0.075	1.252	0.071	1.256	0.060	1.271	0.072	1.104	0.069	1.116	0.059	1.151
116	1.116	0.053	1.402	0.060	1.391	0.149	1.260	0.052	1.295	0.059	1.270	0.141	0.966
117	1.125	0.057	0.964	0.064	0.957	0.126	0.895	0.056	0.851	0.063	0.830	0.118	0.652
118	1.000	0.000	1.250	0.000	1.251	0.050	1.188	0.000	1.250	0.000	1.252	0.049	1.090
119	1.182	0.080	1.109	0.082	1.107	0.090	1.097	0.077	0.951	0.079	0.945	0.086	0.921
120	1.108	0.050	1.209	0.052	1.206	0.073	1.180	0.049	1.110	0.051	1.102	0.071	1.035

Table A.12: Cohen-Poisson model - estimation results for Russian texts

		1-displaced CP						**size-biased CP**			
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\alpha}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$
1	1.247	0.302	1.134	0.147	1.079	0.190	1.094	∅	1.256	∅	1.204
2	1.013	0.008	1.364	∅	1.318	∅	1.337	∅	1.367	0.163	1.301
3	1.189	0.245	1.275	0.176	1.253	0.223	1.267	∅	1.368	∅	1.401
4	1.079	0.085	1.123	0.175	1.154	0.138	1.143	∅	1.155	∅	1.221
5	0.840	∅	1.020	∅	1.106	∅	1.088	0.395	0.938	0.136	1.057
6	1.046	0.037	1.217	0.247	1.289	0.205	1.277	∅	1.231	∅	1.398
7	1.151	0.172	0.891	0.331	0.947	0.293	0.935	∅	0.958	∅	1.109
8	1.528	0.723	1.613	0.436	1.540	0.578	1.570	∅	1.885	∅	1.970
9	1.175	0.236	1.419	0.290	1.435	0.268	1.429	∅	1.505	∅	1.594
10	1.046	0.057	1.215	0.099	1.230	0.105	1.232	∅	1.236	∅	1.292
11	1.016	0.006	1.083	0.100	1.116	0.069	1.106	∅	1.085	∅	1.144
12	1.152	0.201	1.581	0.208	1.585	0.241	1.594	∅	1.649	∅	1.738
13	1.253	0.335	1.568	0.275	1.545	0.230	1.534	∅	1.686	∅	1.672
14	1.035	0.028	1.425	∅	1.354	∅	1.358	∅	1.435	0.356	1.275

Table A.12: Cohen-Poisson model - estimation results for Russian texts

		1-displaced CP						**size-biased CP**			
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\alpha}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$
15	1.164	0.188	1.069	0.232	1.086	0.255	1.094	∅	1.143	∅	1.247
16	1.004	∅	1.074	0.231	1.155	0.171	1.137	0.002	1.074	∅	1.236
17	1.047	0.043	1.380	0.225	1.439	0.179	1.427	∅	1.395	∅	1.532
18	0.816	∅	1.141	∅	1.191	∅	1.182	0.438	1.050	0.338	1.101
19	0.786	∅	1.058	∅	1.142	∅	1.121	0.484	0.955	0.302	1.049
20	1.068	0.076	1.347	0.197	1.389	0.206	1.391	∅	1.374	∅	1.514
21	1.181	0.225	1.227	0.198	1.218	0.212	1.222	∅	1.314	∅	1.349
22	1.190	0.267	1.451	0.118	1.403	0.158	1.415	∅	1.548	∅	1.507
23	1.108	0.139	0.999	0.228	1.030	0.203	1.022	∅	1.052	∅	1.139
24	1.077	0.100	1.161	0.096	1.160	0.097	1.160	∅	1.199	∅	1.215
25	1.246	0.321	1.256	0.102	1.180	0.164	1.201	∅	1.383	∅	1.296
26	1.062	0.073	0.872	0.104	0.883	0.095	0.880	∅	0.899	∅	0.929
27	1.133	0.172	1.219	0.126	1.202	0.136	1.206	∅	1.284	∅	1.283
28	1.150	0.198	1.218	0.135	1.196	0.157	1.203	∅	1.293	∅	1.294
29	1.128	0.169	1.230	0.152	1.225	0.167	1.229	∅	1.294	∅	1.327
30	1.163	0.224	1.395	0.038	1.332	0.101	1.353	∅	1.477	∅	1.411
31	1.311	0.496	1.802	0.184	1.716	0.262	1.734	∅	1.962	∅	1.889
32	1.092	0.136	1.596	0.153	1.601	0.158	1.602	∅	1.641	∅	1.693
33	1.184	0.275	1.611	0.067	1.546	0.118	1.561	∅	1.704	∅	1.628
34	1.161	0.256	1.764	0.181	1.744	0.206	1.750	∅	1.845	∅	1.868
35	1.104	0.160	1.667	0.170	1.670	0.167	1.669	∅	1.719	∅	1.765
36	1.143	0.227	1.743	0.199	1.737	0.232	1.745	∅	1.815	∅	1.880
37	1.046	0.063	1.432	0.003	1.412	0.023	1.419	∅	1.454	∅	1.432
38	1.249	0.355	1.521	0.147	1.456	0.209	1.474	∅	1.649	∅	1.598
39	1.047	0.067	1.538	0.095	1.546	0.073	1.540	∅	1.560	∅	1.581
40	1.126	0.191	1.670	0.171	1.665	0.197	1.671	∅	1.732	∅	1.786
41	1.343	0.533	1.755	0.210	1.665	0.299	1.687	∅	1.933	∅	1.867
42	1.125	0.188	1.612	0.115	1.589	0.126	1.592	∅	1.674	∅	1.664
43	1.217	0.341	1.744	0.115	1.678	0.160	1.690	∅	1.854	∅	1.781
44	1.170	0.269	1.739	0.180	1.714	0.216	1.723	∅	1.825	∅	1.849
45	1.220	0.324	1.589	0.213	1.556	0.248	1.566	∅	1.702	∅	1.715
46	1.213	0.317	1.607	0.184	1.567	0.224	1.578	∅	1.716	∅	1.711
47	1.229	0.359	1.739	0.198	1.695	0.252	1.708	∅	1.856	∅	1.856
48	1.229	0.345	1.645	0.142	1.584	0.191	1.597	∅	1.762	∅	1.709
49	1.083	0.118	1.460	0.125	1.462	0.116	1.460	∅	1.502	∅	1.526
50	1.172	0.274	1.764	0.151	1.729	0.191	1.739	∅	1.851	∅	1.849
51	1.057	0.095	1.859	0.204	1.889	0.185	1.885	∅	1.887	∅	1.988
52	1.169	0.255	1.623	0.115	1.581	0.161	1.594	∅	1.709	∅	1.687
53	1.150	0.233	1.693	0.163	1.673	0.194	1.681	∅	1.768	∅	1.793
54	1.027	0.035	1.432	0.052	1.437	0.044	1.435	∅	1.443	∅	1.459
55	1.170	0.259	1.659	0.176	1.635	0.203	1.642	∅	1.745	∅	1.761
56	1.049	0.070	1.555	0.136	1.576	0.135	1.576	∅	1.578	∅	1.653
57	1.104	0.155	1.563	0.132	1.556	0.144	1.560	∅	1.615	∅	1.643
58	1.154	0.237	1.680	0.177	1.663	0.200	1.669	∅	1.758	∅	1.785
59	1.252	0.368	1.579	0.226	1.536	0.247	1.541	∅	1.709	∅	1.690
60	1.338	0.512	1.693	0.190	1.601	0.285	1.625	∅	1.870	∅	1.798
61	0.802	∅	1.062	∅	1.114	∅	1.105	0.463	0.964	0.359	1.018
62	0.933	∅	1.129	∅	1.092	∅	1.110	0.186	1.093	0.315	1.035

Table A.12: Cohen-Poisson model - estimation results for Russian texts

		1-displaced CP						**size-biased CP**			
No.	d	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\alpha}_{\mathrm{FF}}$	$\hat{\theta}_{\mathrm{FF}}$	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\theta}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\theta}_{\mathrm{ML}}$
63	1.087	0.114	1.094	0.170	1.114	0.167	1.113	∅	1.138	∅	1.209
64	1.064	0.081	1.015	0.140	1.036	0.127	1.032	∅	1.045	∅	1.102
65	0.921	∅	1.019	∅	1.058	∅	1.051	0.207	0.978	0.062	1.038
66	0.869	∅	1.186	∅	1.226	∅	1.219	0.339	1.118	0.235	1.165
67	1.226	0.313	1.408	0.302	1.407	0.335	1.416	∅	1.525	∅	1.628
68	1.408	0.482	0.938	0.341	0.890	0.403	0.909	∅	1.157	∅	1.166
69	1.082	0.098	1.169	∅	1.125	0.012	1.138	∅	1.206	∅	1.144
70	0.761	∅	0.832	∅	0.906	∅	0.881	0.519	0.719	0.404	0.785
71	0.893	∅	1.022	0.052	1.098	∅	1.077	0.273	0.968	0.014	1.074
72	0.905	∅	1.013	0.103	1.100	0.041	1.078	0.241	0.966	∅	1.100
73	0.861	∅	1.325	0.143	1.442	0.055	1.416	0.355	1.256	∅	1.447
74	0.866	∅	0.928	∅	0.965	∅	0.958	0.329	0.861	0.237	0.904
75	0.909	∅	0.949	∅	0.961	∅	0.959	0.244	0.901	0.232	0.906
76	0.891	∅	0.897	∅	0.943	∅	0.932	0.271	0.843	0.134	0.902
77	0.906	∅	0.854	∅	0.876	∅	0.871	0.252	0.804	0.201	0.827
78	1.122	0.161	1.339	0.192	1.350	0.190	1.349	∅	1.398	∅	1.462
79	1.139	0.177	1.241	0.121	1.223	0.160	1.235	∅	1.308	∅	1.328
80	1.094	0.109	1.195	0.139	1.206	0.143	1.208	∅	1.236	∅	1.290
81	0.883	∅	0.988	∅	1.027	∅	1.019	0.288	0.930	0.179	0.978
82	1.014	0.016	1.016	0.050	1.028	0.044	1.026	∅	1.022	∅	1.049
83	1.168	0.214	1.278	0.255	1.293	0.262	1.295	∅	1.359	∅	1.456
84	1.143	0.182	1.053	0.122	1.032	0.151	1.041	∅	1.124	∅	1.126
85	0.808	∅	1.040	∅	1.141	∅	1.116	0.443	0.946	0.163	1.080
86	1.079	0.099	1.073	0.103	1.075	0.117	1.080	∅	1.111	∅	1.145
87	0.916	∅	1.052	∅	1.091	∅	1.087	0.221	1.008	0.065	1.072
88	0.947	∅	0.899	0.027	0.937	0.004	0.928	0.143	0.871	∅	0.930
89	0.952	∅	1.158	0.098	1.219	0.048	1.202	0.139	1.132	∅	1.229
90	1.055	0.070	1.247	0.103	1.260	0.119	1.265	∅	1.273	∅	1.333
91	1.069	0.092	1.225	0.121	1.235	0.122	1.236	∅	1.259	∅	1.306
92	1.071	0.095	1.175	0.110	1.180	0.111	1.181	∅	1.210	∅	1.244
93	1.198	0.284	1.509	0.104	1.451	0.144	1.462	∅	1.610	∅	1.546
94	1.111	0.146	1.163	0.105	1.149	0.117	1.153	∅	1.219	∅	1.219
95	1.164	0.220	1.262	0.149	1.237	0.167	1.243	∅	1.345	∅	1.341
96	1.242	0.357	1.592	0.245	1.558	0.264	1.563	∅	1.717	∅	1.722
97	1.125	0.180	1.576	0.190	1.580	0.202	1.583	∅	1.637	∅	1.702
98	1.034	0.044	1.143	0.071	1.152	0.070	1.152	∅	1.159	∅	1.190
99	1.210	0.281	1.258	0.220	1.237	0.245	1.245	∅	1.367	∅	1.393
100	1.025	0.031	1.266	∅	1.245	∅	1.250	∅	1.277	0.031	1.244
101	1.082	0.111	1.288	0.150	1.302	0.140	1.298	∅	1.329	∅	1.380
102	1.148	0.212	1.481	0.184	1.473	0.205	1.479	∅	1.556	∅	1.600
103	1.053	0.072	1.220	0.124	1.238	0.114	1.235	∅	1.246	∅	1.300
104	1.220	0.293	1.253	0.268	1.245	0.278	1.248	∅	1.368	∅	1.420
105	1.159	0.209	1.146	0.107	1.109	0.130	1.117	∅	1.227	∅	1.190
106	0.986	∅	1.306	0.060	1.334	0.039	1.327	0.040	1.299	∅	1.349
107	1.154	0.214	1.407	0.154	1.387	0.159	1.388	∅	1.485	∅	1.481
108	0.935	∅	1.346	0.117	1.418	0.065	1.403	0.182	1.313	∅	1.440
109	1.019	0.023	1.282	0.088	1.305	0.073	1.300	∅	1.290	∅	1.341
110	1.147	0.200	1.302	0.109	1.270	0.136	1.279	∅	1.376	∅	1.358

Table A.12: Cohen-Poisson model - estimation results for Russian texts

		1-displaced CP						**size-biased CP**			
No.	d	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$	$\hat{\alpha}_{\text{FF}}$	$\hat{\theta}_{\text{FF}}$	$\hat{\alpha}_{\text{MM}}$	$\hat{\theta}_{\text{MM}}$	$\hat{\alpha}_{\text{ML}}$	$\hat{\theta}_{\text{ML}}$
111	1.073	0.100	1.302	0.061	1.288	0.067	1.290	∅	1.339	∅	1.328
112	1.218	0.284	1.191	0.192	1.160	0.227	1.171	∅	1.304	∅	1.307
113	1.090	0.122	1.360	0.138	1.366	0.134	1.364	∅	1.404	∅	1.442
114	1.065	0.088	1.243	0.102	1.248	0.102	1.248	∅	1.275	∅	1.305
115	1.168	0.232	1.432	0.055	1.372	0.085	1.382	∅	1.516	∅	1.430
116	1.116	0.171	1.537	0.228	1.555	0.219	1.553	∅	1.595	∅	1.683
117	1.125	0.162	1.083	0.220	1.103	0.206	1.099	∅	1.146	∅	1.219
118	1.000	∅	1.250	0.087	1.281	0.072	1.276	0.002	1.250	∅	1.316
119	1.182	0.246	1.293	0.105	1.244	0.135	1.254	∅	1.386	∅	1.332
120	1.108	0.150	1.325	0.088	1.304	0.107	1.310	∅	1.380	∅	1.371

Appendix B

Generating Functions and Moments

This appendix summarizes some basic statistical theory on generating functions and theoretical moments required in this thesis and is based on Casella and Berger (2002, Chap.2) and Kendall and Stuart (1958, Chap.3).

Let X be a nonnegative discrete random variable which takes values $x \in \mathbb{N}_0$ with *probability mass function (pmf)* given by $\pi_x = P(X = x)$. Another function associated with a random variable X is its *cumulative distribution function (cdf)*, denoted by $F_X(x)$ and defined as $F_X(x) = P(X \leq x)$ for all $x \in \mathbb{N}_0$. Throughout this work we consider only proper distributions, i.e. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$. Also, we denote here random variables by uppercase letters, while lowercase letters denote their realizations.

Definition B.1 *Let X be a discrete random variable defined over nonnegative integers. The probability generating function (pgf) of X, denoted by $G_X(t)$, is given by polynomial*

$$G_X(t) = \pi_0 + \pi_1 t + \pi_2 t^2 + \ldots = \sum_{i=0}^{\infty} t^i \pi_i = E(t^X). \tag{B.1}$$

The subscript of $G_X(t)$ may be omitted if the random variable is clear from the context, then we simply write $G(t)$. The pmf of the random variable X, as well as its moments, can be derived directly from the pgf. The uniqueness of a polynomial expansion implies that the pgf in fact defines the single probabilities which can be calculated as

$$\pi_i = P(X = i) = \frac{1}{i!} \frac{d^i G_X(t)}{dt^i}\bigg|_{t=0}, \quad i = 0, 1, \ldots \tag{B.2}$$

The *k-th factorial moment*, if it exists, is given by

$$\mu_{(k)} = \mathrm{E}[X(X-1)\ldots(X-k+1)] = \frac{d^k G_X(t)}{dt^k}\bigg|_{t=1} = G_X^{(k)}(1)\,, \quad k = 1, 2, \ldots \tag{B.3}$$

The pgf of X is also closely related to its moment generating function.

Definition B.2 *Let X be a discrete random variable. The moment generating function (mgf) of X, denoted by $M_X(t)$, is defined as*

$$M_X(t) = E(e^{tX}) = \sum_{i=0}^{\infty} e^{ti}\pi_i = G_X(e^t)\,, \tag{B.4}$$

provided that the expectation exists (i.e. is finite) for real values of t. If the expectation does not exist, the mgf does not exist.

If the mgf of X exists, then it can be used to find all raw moments of the distribution. The *k-th raw moment* (or *k-th moment about zero*) is generated by

$$\mu'_k = \mathrm{E}(X^k) = \left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = M_X^{(k)}(0)\,,\quad k = 1, 2, \ldots \tag{B.5}$$

Notice that the k-th raw moment can also arise as k-th derivative of $G_X(e^t)$ evaluated at $t = 0$, since $M_X(t) = G_X(e^t)$. Hence, the first raw moment, called the *mean* of X, is obtained as $\mu = \mathrm{E}(X) = G'(1)$. If mgf does not exist *the characteristic function (cf) of X*, defined by

$$\varphi_X(t) = \mathrm{E}(e^{itX})\,,$$

where $i = \sqrt{-1}$, can be useful. The characteristic function has a great theoretical importance, since it always exists and is unique for all distributions. Like mgf, $\varphi_X(t)$ can be used to generate the raw moments by

$$\mu'_k = (-i)^k \left.\frac{d^k\varphi_X(t)}{dt^k}\right|_{t=0}\,,\qquad k = 1, 2, \ldots, n\,, \tag{B.6}$$

provided that the distribution has finite moments μ'_k up to order n. Apart from pgf and mgf, other generating functions exist. Their relationship to the pgf is presented in Table B.1. Generally, they are less useful than pgf, but in some situations they can simplify calculations. Using cumulant generating function we can derive the *k-th cumulant* of X as $\kappa_k = K_X^{(k)}(0)$. The factorial moment generating function enables

Table B.1: Relationships between generating functions of a random variable X

Generating function	Abbreviation	Relation
probability generating function	pgf	$G_X(t)$
characteristic function	cf	$\varphi_X(t) = G_X(e^{it})$
moment generating function	mgf	$M_X(t) = G_X(e^t)$
factorial moment generating function	fmgf	$G_X(1+t)$
cumulant generating function	cgf	$K_X(t) = \log M_X(t)$

us to express probabilities of discrete distribution in terms of factorial moments. As a consequence, we have (cf. Johnson et al., 1992, p. 50)

$$\pi_i = \sum_{k \geq i} (-1)^{i+k} \binom{k}{i} \frac{\mu_{(k)}}{k!} \,. \tag{B.7}$$

Furthermore, the k-th factorial moment can be calculated directly from the raw moments using the Stirling numbers of the first kind, $s(k,i)$, as follows (cf. Johnson et al., 1992, 43f.) :

$$\mu_{(k)} = \sum_{i=0}^{k} s(k,i)\mu'_i \,. \tag{B.8}$$

For the inverse calculation we use the Stirling numbers of the second kind, $S(k,i)$, as follows :

$$\mu'_k = \sum_{i=0}^{k} S(k,i)\mu_{(i)}. \tag{B.9}$$

Tables B.2 and B.3 give values for both $s(k,i)$ and $S(k,i)$ for $1 \leq k,i \leq 7$. The following holds: $s(0,0) = S(0,0) = 1$ and $s(k,0) = S(k,0) = 0\,, k > 0$. Both kinds of Stirling numbers are non-zero for $i = 1, 2, \ldots, k,\ k > 0$. For a given k, or a given i, the sign of $s(k,i)$ alternate. On the contrary, $S(k,i)$ are always positive.

Table B.2: Stirling numbers of the first kind $s(k,i)$

	i						
k	1	2	3	4	5	6	7
1	1						
2	-1	1					
3	2	-3	1				
4	-6	11	-6	1			
5	24	-50	35	-10	1		
6	-120	274	-225	85	-15	1	
7	720	-1764	1624	-735	175	-21	1

Useful properties and extensive tables of both sets of Stirling numbers can be found in Abramowitz and Stegun (1965, p. 824). Also, both $S(k,i)$ and $s(k,i)$ can be computed directly as

$$S(k,i) = \frac{1}{i!} \sum_{j=0}^{i} (-1)^{i-j} \binom{i}{j} j^k \,, \tag{B.10}$$

$$s(k,i) = \sum_{j=0}^{k-i} (-1)^j \binom{k-1+j}{k-i+j} \binom{2k-i}{k-i-j} S(k-i+j, j) \,. \tag{B.11}$$

Table B.3: Stirling numbers of the second kind $S(k,i)$

	i						
k	1	2	3	4	5	6	7
1	1						
2	1	1					
3	1	3	1				
4	1	7	6	1			
5	1	15	25	10	1		
6	1	31	90	65	15	1	
7	1	63	301	350	140	21	1

In applications the moments of maximally fourth order are of interest. In particular, we have

$$\begin{aligned} \mu'_1 &= \mu_{(1)} = \mu\,, & \mu'_2 &= \mu_{(2)} + \mu\,, \\ \mu'_3 &= \mu_{(3)} + 3\mu_{(2)} + \mu\,, & \mu'_4 &= \mu_{(4)} + 6\mu_{(3)} + 7\mu_{(2)} + \mu\,. \end{aligned} \tag{B.12}$$

The *k-th central moment* can further be calculated from the raw moments using the formula

$$\mu_k = \mathrm{E}((X-\mu)^k) = \sum_{i=0}^{k} \binom{k}{i} \mu'_{k-i}(-\mu)^i\,, \quad k = 1, 2\ldots \tag{B.13}$$

Inverse calculation is less convenient. First four moments are given by

$$\begin{aligned} \mu_1 &= 0\,, & \mu_3 &= \mu'_3 - 3\mu'_2\mu + 2\mu^3\,, \\ \mu_2 &= \mu'_2 - \mu^2\,, & \mu_4 &= \mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4\,. \end{aligned} \tag{B.14}$$

The first central moment μ_1 is always zero, since by definition $\mu'_0 = 1$. The *variance* of a random variable X is its second central moment, $\mu_2 = \mathrm{var}(X)$. The positive square root of $\mathrm{var}(X)$ is the *standard deviation* of X.

Probability generating functions can also be used to find the *convolution* of the two distributions. If X_1 and X_2 are two independent random variables with pgf $G_{X_1}(t)$ and $G_{X_2}(t)$, respectively, then the distribution of their sum $X = X_1 + X_2$ has the pgf given by

$$G_X(t) = \mathrm{E}(t^X) = \mathrm{E}(t^{X_1+X_2}) = \mathrm{E}(t^{X_1}t^{X_2}) = \mathrm{E}(t^{X_1})\mathrm{E}(t^{X_2}) = G_{X_1}(t)G_{X_2}(t)\,. \tag{B.15}$$

Likewise, mgf is

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}(e^{tX_1}e^{tX_2}) = M_{X_1}(t)M_{X_2}(t)\,, \tag{B.16}$$

and cf is

$$\varphi_X(t) = \mathrm{E}(e^{itX}) = \mathrm{E}(e^{itX_1}e^{itX_2}) = \varphi_{X_1}(t)\varphi_{X_2}(t)\,. \tag{B.17}$$

Even more generally, if $X_1, X_2, \dots, X_n$ are mutually independent random variables with pgf's $G_{X_1}(t), G_{X_2}(t), \dots, G_{X_n}(t)$, respectively, then the pgf, mgf and cf of the random variable $X = \sum_{i=1}^{n} X_i$ are given by

$$G_X(t) = \prod_{i=1}^{n} G_{X_i}(t)\,, \quad M_X(t) = \prod_{i=1}^{n} M_{X_i}(t)\,, \quad \varphi_X(t) = \prod_{i=1}^{n} \varphi_{X_i}(t)\,. \tag{B.18}$$

Appendix C

Properties of Estimators

This appendix deals with basic properties of estimators as well as criteria for making decisions which estimator should be taken as the most appropriate one. It is based on the theory given by Casella and Berger (2002).

Let $X_1, \ldots, X_n$ be a sample from a population with pmf $\pi_{x|\theta} = P_\theta(X = x)$ where θ is an unknown parameter. Any function $T = T(X_1, \ldots, X_n)$ of a sample is a point estimator of θ. We define the quality criteria for the function T as follows:

Definition C.1 *The mean square error (MSE) of an estimator T is defined by*

$$MSE_\theta(T) = E_\theta((T - \theta)^2) = var_\theta(T) + Bias^2_\theta(T)\,,$$

where its bias is given by

$$Bias_\theta(T) = E_\theta(T) - \theta\,.$$

The estimator T of a parameter θ is called *unbiased estimator*, if $\mathrm{E}_\theta(T) = \theta$, whereas it is called *asymptotically unbiased* if $\lim\limits_{n\to\infty} \mathrm{E}_\theta(T) = \theta$. Clearly, for an unbiased estimator $MSE_\theta(T) = \mathrm{var}_\theta(T)$ holds.

Definition C.2 *A consistent estimator T is one for which the following holds*

$$\lim_{n\to\infty} P(|T - \theta| > \varepsilon) = 0\,, \ \ \textit{for } \forall\varepsilon > 0\,. \tag{C.1}$$

Apparently, if T is an unbiased estimator, then based on Chebyshev's inequality we have $P(|T - \theta| > \varepsilon) \leq \mathrm{var}(T)/\varepsilon^2$. Hence, $\forall\varepsilon > 0$ if $\lim\limits_{n\to\infty} \mathrm{var}(T) = 0$, T is consistent.

Definition C.3 *An estimator T_1 is an efficient estimator of θ, if for any estimator T_2 of θ the following inequality holds:*

$$E((T_1 - \theta)^2) \leq E((T_2 - \theta)^2)\,. \tag{C.2}$$

Clearly, among all unbiased estimators the most efficient is the one with the smallest variance, since $MSE(T_1) \leq MSE(T_2)$ holds iff $\mathrm{var}(T_1) \leq \mathrm{var}(T_2)$. However, this fact is even more general for the class of estimators $\{T : \mathrm{E}_\theta(T) = g(\theta) \neq \theta\}$ with the same expected value, as stated by the following definition.

Definition C.4 *An estimator T^* is said to be the best unbiased estimator of $g(\theta)$ if it satisfies $E_\theta(T^*) = g(\theta)$ for all θ and for any other estimator T with $E_\theta(T) = g(\theta)$, we have $var_\theta(T^*) \leq var_\theta(T)$ for all θ. T^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $g(\theta)$.*

Suppose further that both T and T^* are unbiased estimators of parameter θ, where T is better than T^*. Then, any weighted average $W_\alpha(T, T^*) = \alpha T + (1-\alpha)T^*$, for $\forall\alpha$, is also an unbiased estimator of θ. The question is, whether T is also a better estimator than any $W_\alpha(T, T^*)$. The following Cramér-Rao theorem gives answer to this question. It states that for any differentiable function $g(\theta)$, if the variability bound of its unbiased estimator is achieved, then we did the best we could. Thus, any unbiased estimator getting at this lower bound is the best unbiased estimator of $g(\theta)$. A UMVUE may not necessarily reach it.

Theorem C.1 *Let $E_\theta(T) = g(\theta)$ and $\mathcal{I}(\theta)$ be the Fisher information for θ based on X. Then we have*

$$var_\theta(T) \geq \frac{(g'(\theta))^2}{\mathcal{I}(\theta)}\,. \tag{C.3}$$

In particular, if $E_\theta(T) = \theta$, then

$$var_\theta(T) \geq \frac{1}{\mathcal{I}(\theta)}\,. \tag{C.4}$$

The quantity $(g'(\theta))^2/\mathcal{I}(\theta)$ is called the Cramér-Rao lower bound (CRLB). Notice that CRLB depends only on the function $g(\theta)$ and the underlying pmf $\pi_{x|\theta}$. As the information increases, we have less uncertainty about θ and smaller bound on the variance of the best unbiased estimator.

Appendix D

Software

This appendix summarizes the names of the programs we used for analyzing word length frequency distributions. The programs documented here are written in statistical software `R`, `version 2.8.0`. The source code is available as zip file. Preprocessing is done by a special software, developed in the programming language `PERL`, which takes a text as an input and outputs various statistical measures such as mean, variance, higher moments, etc., as well as word length frequency distribution of the text. Data sets containing Slovenian and Russian texts are saved as `dataSlo.R` and `dataRus.R`, respectively. Tables below summarize the functions used for the calculations in `R`.

Table D.1: R Functions used

Function Name	Distribution	Description
`recursion1DSP.R`	*1-displaced Sing-Poisson*	calculates probabilities
`recursionSBSP.R`	*size-biased Sing-Poisson*	calculates probabilities
`generate1DSP.R`	*1-displaced Sing-Poisson*	generates random variables
`generateSBSP.R`	*size-biased Sing-Poisson*	generates random variables
`estimation1dSP.R`	*1-displaced Sing-Poisson*	parameter estimation
`estimationSBSP.R`	*size-biased Sing-Poisson*	parameter estimation
`recursion1DHP.R`	*1-displaced Hyper-Poisson*	calculates probabilities
`recursionSBHP.R`	*size-biased Hyper-Poisson*	calculates probabilities
`generate1DHP.R`	*1-displaced Hyper-Poisson*	generates random variables
`generateSBHP.R`	*size-biased Hyper-Poisson*	generates random variables
`estimation1dHP.R`	*1-displaced Hyper-Poisson*	parameter estimation
`estimationSBHP.R`	*size-biased Hyper-Poisson*	parameter estimation
`recursion1DGP.R`	*1-displaced generalized Poisson*	calculates probabilities
`recursionSBGP.R`	*size-biased generalized Poisson*	calculates probabilities
`generate1DGP.R`	*1-displaced generalized Poisson*	generates random variables
`generateSBGP.R`	*size-biased generalized Poisson*	generates random variables
`estimation1dGP.R`	*1-displaced generalized Poisson*	parameter estimation
`estimationSBGP.R`	*size-biased generalized Poisson*	parameter estimation

Table D.2: R Functions used

Function Name	Distribution	Description
`recursion1DCP.R`	*1-displaced Cohen-Poisson*	calculates probabilities
`recursionSBCP.R`	*size-biased Cohen-Poisson*	calculates probabilities
`generate1DCP.R`	*1-displaced Cohen-Poisson*	generates random variables
`generateSBCP.R`	*size-biased Cohen-Poisson*	generates random variables
`estimation1dCP.R`	*1-displaced Cohen-Poisson*	parameter estimation
`estimationSBCP.R`	*size-biased Cohen-Poisson*	parameter estimation

Bibliography

Ablamowicz, R., and Fauser, B. (2011). *CLIFFORD/Bigebra, A Maple Package for Clifford (Co)Algebra Computations.* (©1996-2011, RA&BF)

Abramowitz, M., and Stegun, I. A. (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications, Inc.

Akman, O., Gamage, J., Jannot, J., Juliano, S., Thurman, A., and Whitman, D. (2007). A simple test for detection of length-biased sampling. *JP Journal of Biostatistics*, *1*(2), 189 -195.

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, *2*, 1 -10.

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika*, *13*, 287 - 298.

Altmann, G. (2005). Diversification processes. In Köhler, R., Altmann, G., and Piotrowski, R. G. (Eds.), *Quantitative Linguistics. An International Handbook.* Berlin: Walter de Gruyter, 646 -658.

Altmann, G., and Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. *Glottometrika*, *15*, 166 -180.

Altmann, G., Best, K.-H., and Wimmer, G. (1997). Wortlänge in romanischen Sprachen. In Gather, A. and Werner, H. (Eds.), *Semiotische Prozesse und Natürliche Sprache. Festschrift für Udo L. Figge zum 60. Geburtstag.* Stuttgart: Steiner, 1 -13.

Altmann, G., Erat, E., and Hřebíček, L. (1996). Word length distribution in Turkish texts. *Glottometrika*, *15*, 195 -204.

Altmann, G., and Hammerl, R. (1989). *Diskrete Wahrscheinlichkeitsverteilungen I: Vol. 41 Quantitative Linguistics.* Bochum: Universitätsverlag Dr. N. Brockmeyer.

Altmann, G., and Zörnig, P. (1992). *Diskrete Wahrscheinlichkeitsverteilungen II: Vol. 47 Quantitative Linguistics.* Bochum: Universitätsverlag Dr. N. Brockmeyer.

Anscombe, F. J. (1950). Sampling theory for the negative binomial and logarithmic series distributions. *Biometrika*, *37*, 358-382.

Antić, G., Grzybek, P., and Stadlober, E. (2005). Mathematical aspects and modifications of Fucks' generalized Poisson distribution (GPD). In Köhler, R., Altmann, G., and Piotrowski, R. G. (Eds.), *Quantitative Linguistics. An International Handbook.* Berlin: Walter de Gruyter, 158-180.

Antić, G., Kelih, E., and Grzybek, P. (2006a). Zero-syllable words in determining word length. In Grzybek, P. (Ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 117-156.

Antić, G., Stadlober, E., Grzybek, P., and Kelih, E. (2006b). Word length and frequency distributions in different text genres. In Spiliopoulou, M., Kruse, R., Nürnberger, A., Borgelt, C., and Gaul, W. (Eds.), *From Data and Information Analysis to Knowledge Engineering.* Heidelberg/Berlin: Springer, 310-317.

Bardwell, G. E., and Crow, E. L. (1964). A two-parameter family of hyper-Poisson distributions. *Journal of the American Statistical Association*, *59*, 133-141.

Bartkowiakowa, A., and Gleichgewicht, B. (1964). Zastosowanie dwuparametrowych rozkladów Fucksa do opisu dlugości sylabicznej wyrazów w różnych utworach prozaicznych autorów polskich. *Zastosowania Matematyki*, *7*, 345-352.

Bartkowiakowa, A., and Gleichgewicht, B. (1965). O rozkladach dlugości sylabicznej wyrazów w różnych tekstach. In Mayenowa, M. R. (Ed.), *Poetyka i Matematyka.* Warszwawa, 164-173.

Best, K.-H. (2001). Wortlängen in Texten gesprochener Sprache. *Göttinger Beiträge zur Sprachwissenschaft*, *6*, 31-42.

Best, K.-H. (2005). Wortlänge. In Köhler, R., Altmann, G., and Piotrowski, R. G. (Eds.), *Quantitative Linguistics. An International Handbook.* Berlin: Walter de Gruyter, 260-273.

Best, K.-H., and Čebanov, S. V. (2001). Biographische Notiz: Sergej Grigor'evič Čebanov (1897–1966). In Best, K.-H. (Ed.), *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt, 281-283.

Best, K.-H., and Zinenko, S. (1998). Wortlängenverteilungen in Briefen A. T. Twardowskis. *Göttingen Beiträge zur Sprachwissenschaft*, *1*, 7-19.

Bühler, H., Fritz, G., and Herlitz, W. (1972). *Linguistik I. Lehr- und Übungsbuch zur Einführung in die Sprachwissenschaft.* Tübingen: Niemeyer.

Böhning, D. (1998). Zero-inflated Poisson models and C.A.MAN: A tutorial collection of evidence. *Biometrical Journal*, *40*(7), 833-843.

Bünting, K. D., and Bergenholtz, H. (1995). *Einführung in die Syntax.* Weinheim: Beltz-Athenäum.

Bokučava, N. V., and Gačečiladze, T. G. (1965). Ob odnom metode izučenija statističeskoj struktury pečatnoj informacii. *Trudy Tbiliskogo Gosudarstvennogo Universiteta*, *103*, 173-180.

Butler, R. W., and Wood, A. T. A. (2002). Laplace approximations for hypergeometric functions with matrix agrument. *The Annals of Statistics*, *30*(4), 1155-1177. (http://projecteuclid.org/euclid.aos/1031689021)

Cameron, C. A., and Trivedi, P. K. (1998). *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

Casella, G., and Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove: Duxbury.

Cercvadze, G. N., Čikoidze, G. B., and Gačečiladze, T. G. (1959). Primenenie matematičeskoj teorii slovoobrazovanija k gruzinskomu jazyku. *Soobščenija Akademii Nauk Gruzinskoj SSR*, *22, 6*, 705-710.

Chebanow, S. G. (1947). On conformity of language structures within the Indo-European family to Poisson's law. *Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS*, *55*(2), 99-102.

Cohen, C. A. (1959). Estimation in the Poisson distribution when sample values of $c + 1$ are sometimes erroneously reported as c. *Annals of the Institute of Statistical Mathematics*, *11*, 189-193.

Cohen, C. A. (1991). *Truncated and Censored Samples. Theory and Applications.* New York: Marcel Dekker, Inc.

Consul, P. C. (1989). *Generalized Poisson distributions. Properties and Applications.* New York: Marcel Dekker, Inc.

Consul, P. C., and Famoye, F. (1988). Maximum likelihood estimation for the generalized Poisson distribution when the sample mean is larger than sample variance. *Comm.Statist.-Theory Methods*, *17*(1), 299-309.

Consul, P. C., and Famoye, F. (2006). *Lagrangian Probability Distributions.* Boston: Birkhäuser.

Consul, P. C., and Jain, G. C. (1973a). A generalization of the Poisson distribution. *Technometrics*, *15*(4), 791-799.

Consul, P. C., and Jain, G. C. (1973b). On some interesting properties of the generalized Poisson distribution. *Biometrische Zeitschrift*, *15*, 495-500.

Consul, P. C., and Shoukri, M. M. (1984). Maximum likelihood estimation for the generalized Poisson distribution. *Comm.Statist.-Theory Methods*, *13*(2), 1533-1574.

Consul, P. C., and Shoukri, M. M. (1985). The generalized Poisson distribution when the sample mean is larger than the sample variance. *Comm.Statist.-Simulation Comput.*, *14*(3), 667-681.

Cressie, N. A. C., and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, *B46*(3), 440-464.

Cressie, N. A. C., and Read, T. R. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* New York: Springer.

Crow, E. L., and Bardwell, G. E. (1965). Estimation of the parameters of the hyper-Poisson distributions. In Patil, G. P. (Ed.), *Classical and Contagious Discrete Distributions (Proc. Internat. Sympos., McGill Univ., Montreal, Canada, 1963).* Calcutta: Statistical Publishing Society, 127-140.

Dietz, E., and Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics* & *Data Analysis*, *34*, 441-459.

Djuraš, G., and Stadlober, E. (2010). Modeling word length frequencies by the Singh-Poisson distribution. In Grzybek, P., Kelih, E., and Mačutek, J. (Eds.), *Text and Language. Structures · Functions · Interrelations · Quantitative · Perspectives.* Wien: Praesens, 37-48.

Famoye, F., and Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*, *4*, 117-130.

Feller, W. (1943). On a general class of contagious distributions. *Ann. Math. Statist.*, *14*, 389-400.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* New York: John Wiley and Sons, Inc.

Fisher, R. A. (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, *6*, 13-25.

Fucks, W. (1955). Theorie der Wortbildung. In Behnke, H., Lietzmann, W., and Süss, W. (Eds.), *Mathematisch-Physikalische Semesterberichte, 4.* Göttingen: Vandenhoeck & Ruprecht, 195-212.

Fucks, W. (1956a). Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen. *Nachrichtentechnische Fachberichte*, *3*, 7-21.

Fucks, W. (1956b). Statistische Verteilungen mit gebundenen Anteilen. *Zeitschrift für Physik*, *145*, 520-533.

Fucks, W. (1956c). Mathematical theory of word formation. In Cherry, C. (Ed.), *Information Theory.* London: Butterworths, 154-170.

Fucks, W. (1957). Gibt es mathematische Gesetze in Sprache und Musik? *Umschau*, 57_2, 33-37.

Gačečiladze, T. G., Cercvadze, G. N., and Čikoidze, G. B. (1961). Ob ε-strukture raspredelenija probelov. *Trudy Instituta Élektroniki, Avtomatiki i Telemechaniki*, *2*, 3-15.

Gačečiladze, T. G., and Cilosani, T. P. (1971). Ob odnom metode izučenija statističeskoj struktury teksta. *Statistika Reči i Avtomatičeskij Analiz Teksta*, 113-133.

Girzig, P. (1997). Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. *Glottometrika*, *16*, 152-162.

Gove, J. H. (2003a). Estimation and applications of size-biased distributions in forestry. In Amaro, A., Reed, D., and Soares, P. (Eds.), *Modelling Forest Systems.* Cambridge: CABI, 201-212.

Gove, J. H. (2003b). Moment and maximum likelihhod estimators for Weibull distributions under length- and area-biased sampling. *Enviromental and Ecological Statistics*, *10*, 455-467.

Grotjahn, R. (1982). Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft*, *1*, 44-75.

Grotjahn, R., and Altmann, G. (1993). Modelling the distribution of word length: Some methodological problems. In Köhler, R. and Rieger, B. (Eds.), *Contributions to Quantitative Linguistics.* Dordrecht: Kluwer, 141-153.

Grzybek, P. (2006). History and methodology of word length studies. In Grzybek, P. (Ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 15-90.

Grzybek, P., Kelih, E., and Stadlober, E. (2005a). Empirische Textsemiotik und quantitative Texttypology. In Bernard, J, Fikfak, J., and Grzybek, P. (Eds.), *Text & Reality.* Ljubljana: Založba ZRC, 95-120.

Grzybek, P., Stadlober, E., Kelih, E., and Antić, G. (2005b). Quantitative text typology: The impact of word length. In Weihs, C. and Gaul, W. (Eds.), *Classification - The Ubiquitous Challenge.* Heidelberg: Springer, 53-64.

Gupta, P. L., Gupta, R. C., and Tripathi, R. C. (1996). Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis*, *23*, 207-218.

Gurland, J. (1957). Some interrelations among compound and generalized distributions. *Biometrika*, *44*, 265-268.

Gurland, J. (1958). A generalized class of contagious distributions. *Biometrics*, *14*, 229-249.

Gurland, J. (1965). A method of estimation for some generalized Poisson distributions. In Patil, G. P. (Ed.), *Classical and Contagious Discrete Distributions (Proc. Internat. Sympos., McGill Univ., Montreal, Canada, 1963).* Calcutta: Statistical Publishing Society, 141-158.

Haight, F. A. (1967). *Handbook of the Poisson distribution.* New York: John Wiley & Sons, Inc.

Hankin, R. (2006). Special functions in R: Introducing the gsl package. *R News*, *6*(4), 24-26.

Joe, H., and Zhu, R. (2005). Generalized Poisson disstribution: The property of mixture of Poisson and comparison with negative binomial disstribution. *Biometrical Journal*, *47*(2), 219-229.

Johnson, N. L., and Kotz, S. (1969). *Discrete Distributions.* New York: John Wiley & Sons, Inc.

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). *Univariate Discrete Distributions* (2nd ed.). New York: John Wiley & Sons, Inc.

Katti, S. K. (1966). Interrelations among generalized distributions and their components. *Biometrics*, *22*, 44-52.

Kelih, E. (2007). Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen. In Kaliuscenko, V., Köhler, R., and Levickij, V. (Eds.), *Problems of Typological and Quantitative Lexicology.* Chernivtsi: Ruta, 91-105.

Kelih, E. (2009a). Zur Homogenität von Graphemhäufigkeiten in Texten: Evidenz aus dem Russischem. In Kelih, E., Levickij, V. V., and Altmann, G. (Eds.), *Methods of Text Analysis.* Chernivtsi: ČNU, 85-105.

Kelih, E. (2009b). Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In Kelih, E., Levickij, V. V., and Altmann, G. (Eds.), *Methods of Text Analysis.* Chernivtsi: ČNU, 106-124.

Kelih, E. (2011). Zum Analytismus und Synthetismus in slawischen Sprachen: Morphologische Wortstruktur in slawischen Sprachen. In Bente Karl, K., Krumbholz, G., and Lazar, M. (Eds.), *Beiträge der Europäischen Slavistischen Linguistik (Polyslav). Band 14.* München/Berlin: Sagner (Die Welt der Slaven. Sammelbände, Sborniki, 43), 99-107.

Kelih, E. (2012). On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts. In Grzybek, P., Naumann, S., and Vulanović, R. (Eds.), *Festschrift für Reinhard Köhler.* Wien: Praesens [In print].

Kelih, E., Antić, G., Grzybek, P., and Stadlober, E. (2005). Classification of author and/or genre? The impact of word length. In Weihs, C. and Gaul, W. (Eds.), *Classification - The Ubiquitous Challenge.* Heidelberg: Springer, 498-505.

Kelih, E., Buk, S., Grzybek, P., and Rovenchak, A. (2009). Project description: Designing and constructing a typologically balanced Ukrainian text database. In Kelih, E., Levickij, V. V., and Altmann, G. (Eds.), *Methods of Text Analysis.* Chernivtsi: ČNU, 125-132.

Kelih, E., and Grzybek, P. (2004). Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte). *Glottometrics*, *8*, 23-41.

Kelih, E., Grzybek, P., and Stadlober, E. (2003). Das Grazer Projekt zu Wortlängen(häufigkeiten). *Glottometrika*, *6*, 94-102.

Kemp, A. W. (2005). Modeling under- and over-dispersion using q-analogues of the Poisson distribution. *American Journal of Mathematical and Management Science*, *25*, 313-342.

Kendall, M. G., and Stuart, A. (1958). *The Advanced Theory of Statistics* (Vol. 1). London: Charles Griffin & Co.

Krámský, J. (1969). *The word as a linguistic unit.* The Hague: Mouton.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14.

Larose, D. T., and Dey, D. K. (1996). Weighted distributions viewed in the context of model selection: A Bayesian perspective. *Test, The Journal of the Spanish Statistical Society*, *5*(1), 227-246.

Lee, A. H., Wang, K., and Yau, K. K. W. (2001). Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal*, *43*(8), 963-975.

Lehfeldt, W. (1999). Akzent. In Jachnow, H. (Ed.), *Handbuch der sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen.* Wiesbaden: Harrassowitz (Slavistische Studienbücher, N.F. 8), 34-48.

Lord, R. D. (1958). Studies in the history of probability and statistics. VIII. De Morgan and the statistical study of literary style. *Biometrika*, *45*, 282.

Maceda, C. E. (1948). On the compound and generalized Poisson distributions. *The Annals of Mathematical Statistics*, *19*, 414-416.

Mačutek, J., Švehlíková, Z., and Cenkerová, Z. (2011). Towards a model for rank-frequency distributions of melodic intervals. *Glottometrics*, *21*, 60-64.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, *9*(214), 237-249.

Mendenhall, T. C. (1901). A mechanical solution of a literary problem. *Popular Science Monthly*, *60*, 97-105.

Moore, D. S. (1984). Measures of lack of fit from tests of chi-squared type. *Journal of Statistical Planning and Inference*, *10*, 151-166.

Nemcová, E., and Altmann, G. (1994). Zur Wortlänge in slowakischen Texten. *Zeitschrift für Empirische Textforschung*, *1*, 40-43.

Neubauer, G., and Djuraš, G. (2008). A generalized Poisson model for underreporting. In Eilers, P. (Ed.), *Proceedings of the 23rd International Workshop on Statistical Modelling, Utrecht, Netherlands, 7-11 July 2008,* 368-373.

Neubauer, G., and Djuraš, G. (2009). A beta-Poisson model for underreporting. In Booth, J. (Ed.), *Proceedings of the 24th International Workshop on Statistical Modelling, Ithaca, NY, 20-24 July 2009,* 255-260.

Neubauer, G., Djuraš, G., and Friedl, H. (2009). *Maximum likelihood for size-estimation: Some results on properties and limitations* (Tech. Rep. No. 4). Graz: Joanneum Research.

Neubauer, G., Djuraš, G., and Friedl, H. (2010). Models for underreporting: A Bernoulli sampling approach for reported counts. *Austrian Journal of Statistics*, *40*(1), 85-92.

Neubauer, G., and Friedl, H. (2006). Modelling sample sizes of frequencies. In Hinde, J., Einbeck, J., and Newell, J. (Eds.), *Proceedings of the 21th International Workshop on Statistical Modelling, Galway, Ireland, 3-7 July 2006,* 401-408.

Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *Annals of Math. Stat.*, *10*, 35-57.

Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics*, *21*, 127-132.

Özmen, İ., and Famoye, F. (2007). Count regression models with an application to zoological data containing structural zeros. *Journal of Data Science*, *5*, 491-502.

Patil, G. P. (2002). Weighted distributions. In El-Shaarawi, A. H. and Piegorsch, W. W. (Eds.), *Encyclopedia of Environmetrics.* Chichester: John Wiley & Sons, 2369-2377.

Patil, G. P., and Rao, R. C. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, *34*, 179-189.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

Piotrowski, R. G., Bektaev, K. B., and Piotrowskaja, A. A. (1985). *Mathematische Linguistik.* Bochum: Studienverlag Dr. N. Brockmeyer.

Puig, P., and Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, *101*(473), 332-340.

Puig, P., and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli*, *13*(2), 544-555.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria. (ISBN 3-900051-07-0)

Rao, R. C. (1965). On discrete distributions arising out of methods of ascertainment. In Patil, G. P. (Ed.), *Classical and Contagious Discrete Distributions (Proc. Internat. Sympos., McGill Univ., Montreal, Canada, 1963).* Calcutta: Statistical Publishing Society, 320-332.

Ridout, M., Demétrio, C. G., and Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference, Cape Town,* 179-192.

Rottmann, O. A. (2006). Aspects of the typology of Slavic languages. In Grzybek, P. (Ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 241-258.

Satterthwaite, F. E. (1942). Generalized Poisson distribution. *Ann. Math. Statist.*, *13*(4), 410-417.

Schwarz, H. R., and Köckler, N. (2006). *Numerische Mathematik.* Wiesbaden: Teubner.

Stadlober, E. (1989). *Sampling from Poisson, binomial and hypergeometric distributions: Ratio of uniforms as a simple and fast alternative* (Bericht No. 303). Graz: Mathematisch-Statistische Sektion, Forschungsgesellschaft Joanneum.

Stadlober, E., and Djuzelic, M. (2006). Multivariate statistical methods in quantitative text analyses. In Grzybek, P. (Ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 259-275.

Strauss, U., Grzybek, P., and Altmann, G. (2006). Word length and word frequency. In Grzybek, P. (Ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 277-294.

Teicher, H. (1960). On the mixture of distributions. *Ann. Math. Statist.*, *31*(1), 55-73.

Uhlířová, L. (1996). How long are words in Czech? *Glottometrika*, *15*, 134-146.

Uhlířová, L. (1997). Word length distributions in Czech: On the generality of linguistic laws and individuality of texts. *Glottometrika*, *16*, 163-173.

Vranić, V. (1965). Statsitičko istraživanje hrvatskosrpskog jezika. *Statistička Revija*, *15*(2-3), 174-185.

Vranić, V., and Matković, V. (1965). Mathematical theory of the syllablic structure of Croato-Serbian. *Rad JAZU (Odjel za Matematičke, Fizičke i Tehničke Nauke; 10)*, *331*, 181-199.

Williams, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, *31*, 356-361.

Williams, C. B. (1956). Studies in the history of probability and statistics: IV. A note on an early statistical study of literary style. *Biometrika*, *43*, 248-256.

Williams, C. B. (1975). Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika*, *62*, 207-212.

Wilson, A. (2006). Word-length distribution in present-day Lower Sorbian newspaper texts. In Grzybek, P. (Ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 319-327.

Wimmer, G., and Altmann, G. (1996). A theory of word length: Some results and generalizations. *Glottometrika*, *15*, 112-133.

Wimmer, G., and Altmann, G. (1999). *Thesaurus of Univariate Discrete Probability Distributions.* Essen: Stamm.

Wimmer, G., and Altmann, G. (2005). Unified derivations of some linguistics laws. In Köhler, R., Altmann, G., and Piotrowski, R. G. (Eds.), *Quantitative Linguistics. An International Handbook.* Berlin: Walter de Gruyter, 791-807.

Wimmer, G., Köhler, R., Grotjahn, R., and Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, *1*(1), 98-106.

Winkelmann, R. (2000). *Econometric Analysis of Count Data.* Berlin: Springer.

Zerzwadse, G., Tschikoidse, G., and Gatschetschiladse, T. (1962). Die Anwendung der mathematischen Theorie der Wortbildung auf die georgische Sprache. *Grundlagenstudien aus Kybernetik und Geisteswissenschaft*, *4*, 110-118.