book, I am making LEXSTATS available under the GNU General Public License. LEXSTATS is a suite of programs written in C including a graphical user interface written in Tcl/Tk. Updates can be obtained from the author by e-mail at baayen@mpi.nl. LEXSTATS is supported for LINUX only. It should run without problems on UNIX platforms, and the individual C-programs will probably run on other platforms as well. All C-programs require input and produce output that is in the *data frame* format of R and Splus, so that the user is not limited to the functionality provided by the graphical user interface. Finally, Appendix D summarizes the frequency distributions of the main data sets analyzed in this book.

I am indebted to Stefan Evert, Estate Khmaladze, Anke Luedeling, Richard Sproat, Arjuna Tuzzi, and especially Kyo Kageura for their careful reading of the manuscript and their detailed comments and suggestions for improvement. Most of all, I am indebted to Fiona Tweedie, with whom I have had the opportunity to collaborate on various issues discussed in this book. Without this wonderful collaboration, the chapter on mixture distributions would not exist. I remember with gratitude my friendship with Rezo Chitashvili, to whom this book is dedicated. It is a pleasure to be able to write that the Yule-Simon model, which he developed, emerges from the present study as an excellent model for word frequency distributions. The idea of writing a book along the present lines was born in the year before his untimely death. Thanks are also due to Antoinette Renouf, who kindly provided the data sets from her large longitudinal corpus of British newspapers, to Stephen Tweedie, who introduced me to the LINUX operating system, and to Jorn Baayen, who has been an excellent LINUX system administrator. And thanks to Tineke for making it all worthwhile.

Chapter 1

Word Frequencies

This chapter introduces two fundamental issues in lexical statistics. The first issue concerns the role of the sample size, the number of words in a text or corpus. The sample size crucially determines a great many measures that have been proposed as characteristic text constants. However, the values of these measures change systematically as a function of the sample size. Similarly, the parameters of many models for word frequency distribution are highly dependent on the sample size. This property sets lexical statistics apart from most other areas in statistics, where an increase in the sample size leads to enhanced accuracy and not to systematic changes in basic measures and parameters.

The second issue concerns the theoretical assumption underlying all theoretical models and tests used in lexical statistics, namely that words occur randomly in texts. This assumption is an obvious simplification that, however, offers the possibility of deriving useful formulae for text characteristics. The crucial question, however, is to what extent this simplifying assumption affects the reliability of these formulae when applied to actual texts and corpora.

Section 1.1 illustrates these two issues by means of an exploratory investigation of word frequencies in Lewis Carroll's *Alice's Adventures in Wonderland*, henceforth *Alice in Wonderland*. Although this is a small book with only 26505 words, it is large enough to reveal the kind of phenomena that emerge, often more strongly, in larger novels and text corpora. Section 1.2 introduces the fundamental concept of the frequency spectrum. Sections 1.3–1.5 review Zipf's rank-frequency model and the lognormal model, as well as a series of statistics that have been proposed as characteristic size-invariant textual constants.

A first objective of this chapter is to show that these statistics and many model parameters are seriously affected by changes in sample size as well as by the non-random organization of discourse. Another equally important objective is to familiarize the reader with some fundamental concepts and notational conventions.

1.1 Introduction

Consider the first sentence of *Alice in Wonderland*:

> Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

This sentence contains 57 instances of words, some of which are used more than once. The function words *of, or* and *the* occur three times, *Alice* and nine other words occur twice, the remaining 28 words occur only once. In all, this sentence with 57 word instances or **word tokens** contains 41 distinct letter strings or **word types.**[1]

Obviously, these frequency counts are strictly bound to this particular sentence. In larger fragments of *Alice in Wonderland*, the frequency counts for these words are quite different. For instance, by the end of the book, Alice has been mentioned 398 times. While *sister* occurs twice in the first sentence just as *Alice* does, it appears only eight times in the complete text. Conversely, the determiner *the*, which appears three times in the first sentence, has a frequency count of 1631 in the novel as a whole. These counts illustrate a simple fact: word frequency depends on sample size. Denoting the **sample size**, the number of word tokens in the sample, by N,

Definition 1.1 N: sample size in word tokens,

I will make the dependency on the sample size of the frequency of the i-th word (ω_i) in a list of word types explicit in my notation:

Definition 1.2 $f(i, N)$: frequency of ω_i in a sample of N tokens.

Table 1.1 presents part of the **word frequency list** of the complete text of *Alice in Wonderland*. For each word ω_i, the frequency $f(i, N)$ is specified.

When we increase the size of our sample, for instance, from the 57 tokens of the first sentence to the 26505 tokens in the complete text of *Alice in Wonderland*, we not only find that the frequencies of the words we have already seen increase, we also encounter new types. The number of different types $V(N)$ we count in a sample of N tokens, the **vocabulary size**, is a non-decreasing function of N.

Definition 1.3 $V(N)$: number of types in a sample of N tokens.

The solid line in panel A of Figure 1.1 plots the development of the vocabulary size $V(N)$ in *Alice in Wonderland* as a function of the sample size N, measured at twenty equally-spaced intervals.

Clearly, the growth curve of the vocabulary size $V(N)$ is not a linear function of N. Initially, the vocabulary size increases quickly, but the rate at which

[1] This is a string-based definition of types and tokens. Alternatively, inflectional variants such as *conversation* and *conversations* can be classified as two tokens of the same type, instead of treating them as tokens of two different types.

Table 1.1: *Part of the word frequency list for* Alice in Wonderland. *i: arbitrary index for the word types; ω_i: the i-th word type; $f(i, 26505)$: the frequency of the i-th word type in the full text of 26505 word tokens.*

i	ω_i	$f(i, 26505)$	i	ω_i	$f(i, 26505)$
1	a	629	23	of	510
2	alice	386	24	on	194
3	alice's	12	25	once	34
4	and	866	26	one	102
5	bank	3	27	or	77
6	beginning	14	28	peeped	3
7	book	7	29	pictures	4
8	but	170	30	reading	3
9	by	57	31	she	540
10	conversation	10	32	sister	8
11	conversations	1	33	sitting	10
12	do	81	34	the	1631
13	get	46	35	thought	74
14	had	177	36	tired	7
15	having	10	37	to	726
16	her	247	38	twice	5
17	in	365	39	use	18
18	into	67	40	very	144
19	is	108	41	was	356
20	it	528	42	what	136
21	no	90	43	without	26
22	nothing	34	44	...	

the vocabulary size increases as we proceed through the text decreases. By the end of the novel the vocabulary growth curve has not flattened out to a horizontal line. A horizontal line would have implied that no new words are added as N increases, which would have indicated that the full set of words judged by Carroll to be appropriate for this kind of story had been used. Instead, it is clear that if the story had continued, more new words would have appeared. Although we can regard *Alice in Wonderland* as the statistical population when we focus on this story as a literary unit, we can equally well view *Alice in Wonderland* as a sample of Carroll's language use. From the latter perspective, the shape of the growth curve $V(N)$ reveals that we have only just begun to sample Carroll's vocabulary.

Suppose that we want to compare *Alice in Wonderland* with *Through the looking-glass and what Alice found there*, henceforth *Through the looking-glass*. The latter is Carroll's second story about Alice. We might hypothesize that Carroll benefited from his experience in writing *Alice in Wonderland*, and that his greater experience as a writer might have lead to a more abundant use of the lexical resources of English. In other words, *Through the looking-glass* might be characterized by the greater vocabulary richness. Comparing $V(N)$ for the two books, we find that *Through the looking-glass* ($V(29028) = 2877$) has

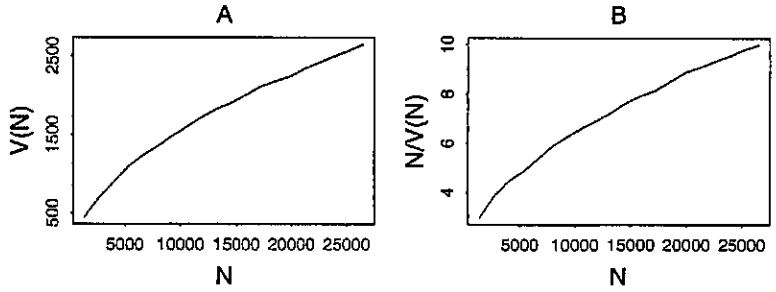

Figure 1.1: *Vocabulary size $V(N)$ (panel A) and mean word frequency $N/V(N)$ (panel B) as a function of sample size N in* Alice in Wonderland, *measured at 20 equally-spaced intervals.*

226 more types than *Alice in Wonderland* ($V(26505) = 2651$). However, since *Alice in Wonderland* is shorter than *Through the looking-glass*, we cannot simply compare their respective vocabulary sizes. If *Alice in Wonderland* had been longer, it would have contained more different words, possibly even more than *Through the looking-glass*.

To adjust for the difference in text size, it seems reasonable at first sight to consider the mean token frequency of the word types, henceforth the **mean word frequency**. A greater vocabulary richness, one would think, should be reflected in a lower mean word frequency $N/V(N)$. For *Alice in Wonderland*, the mean word frequency is 10.00, for *Through the looking-glass*, the mean frequency equals 10.09. These numbers suggest that our hypothesis is wrong, and that Carroll did not exploit the lexical resources of English more fully in his second book.

Interestingly, this conclusion is unjustified. To see this, consider Panel B of Figure 1.1, which plots the mean word frequency $N/V(N)$ for *Alice in Wonderland* at twenty equally-spaced measurement points. The solid line shows the development of the mean through sampling time. Instead of randomly fluctuating around some fixed value, the mean word frequency increases non-linearly with the sample size in a similar way as the vocabulary size $V(N)$. Not only the mean, but also the median changes. For the first six measurement points, the median frequency equals 1, for the remaining measurement points, it equals 2. Apparently, simple statistics such as mean and median do not converge to their population values within the sample. This observation holds not only for a small book such as *Alice in Wonderland*, it generalizes to large novels and even to text corpora with tens of millions of words. Normally, individual sample means fluctuate randomly around the theoretical mean, with larger deviations for smaller samples and smaller deviations for larger samples. Apart from the magnitude of the deviations, there usually is no systematic pattern to the changes in the sample means as the sample size is increased. In the domain of lexical statistics, however, mean frequencies behave surprisingly differently, revealing a non-linear increase with N. In fact, panel B of Figure 1.1 seems a flat contradiction of the fact that the sample mean should become an increasingly accurate estimate of the population mean as the sample size increases. As we shall see, this is due to two factors. One factor is that we are dealing with counts of types instead of with properties of types. A second factor is the large numbers of extremely low-probability words that are present in lexical frequency distributions, distributions that belong to the class of **Large Number of Rare Events** (LNRE) distributions.

Table 1.2: *Sample size N, vocabulary size $V(N)$, mean $N/V(N)$, standard deviation (stdev) and median word frequency for* Through the looking-glass *and* Alice in Wonderland, *as well as for the first 26505 words in* Through the looking-glass.

	N	$V(N)$	$N/V(N)$	stdev	median
Through the looking-glass	29028	2877	10.09	50.91	2
	26505	2731	9.71	47.31	2
Alice in Wonderland	26505	2651	10.00	51.14	2

The dependency of the sample mean on the sample size implies that we have to correct for the difference in sample size before comparing the sample means. Table 1.2 shows that when we compare an equal number of tokens of the two texts, for instance, by selecting the first 26505 words of *Through the looking-glass*, we find that mean frequency for this text has decreased from 10.09 to 9.71, a change in the direction of our original hypothesis. Normally, fluctuations in the value of a mean are due to sampling error and should not be assigned significance. But we have seen that for our lexical data the sample mean increases as a function of the sample size. In our example, the smaller mean frequency observed for $N = 26505$ is not due to sampling error, it is a systematic change in the expected direction. From this point of view, we underestimate the difference in vocabulary richness between the two texts when we use the means of the complete texts. But if we adjust for sample size, we might as well compare the vocabulary sizes directly, instead of focusing on mean frequency. Carroll's second book contains 80 more types among its first 26505 tokens than his first, 3% of the vocabulary size of *Alice in Wonderland*. This suggests that *Through the looking-glass* displays the greater vocabulary richness. In section 3.6, I will introduce a technique for testing whether this difference in vocabulary size is statistically significant.

A second fundamental issue in lexical statistics concerns the non-random use of words in actual texts. However, to construct probabilistic models for word frequency distributions, models that, for instance, yield expressions for $V(N)$ as a function of N, it is convenient to assume that words occur randomly in texts. This is an obvious simplification. In a random rearrangement of all the words of *Alice in Wonderland*, the first 57 words are:

More find likely a somebody a you're lost again was you invent waited a on to time passion so partner about and with panting back-somersault queen as was were the open obliged ask the Alice much a do your as on if face come crab best not rapped gryphon I affair I to it see unlocking low.

Unlike the first sentence of *Alice in Wonderland*, this sequence of words is not semantically coherent. Moreover, sequences of words occur that are ruled out by the rules of syntax (*the Alice, a you're, was were, I to it see*). However, the methodological point at issue here is not whether the randomness assumption is wrong, but to what extent the simplifying assumption of random word use affects the accuracy of theoretical models. Are the effects of non-randomness visible at higher levels of abstraction? Do they introduce significant deviation between theoretically predicted and empirically observed values for statistics such as the vocabulary size $V(N)$? Are effects of non-randomness visibly present in the frequencies of individual words?

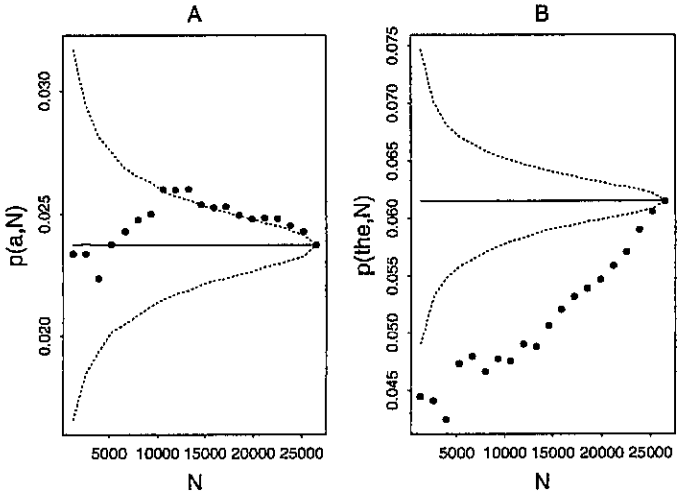

Figure 1.2: *The sample relative frequency of the article* a $p(a, N)$ *(panel A) and the sample relative frequency of the article* the $p(the, N)$ *(panel B) as a function of the sample size* N *in* Alice in Wonderland, *measured at 20 equally-spaced intervals. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on a total of 5000 permutation runs.*

Figure 1.2 plots the **sample relative frequencies** $p(i, N)$ of the definite and indefinite article in *Alice in Wonderland* using large dots.

Definition 1.4 $p(i, N) = \frac{f(i,N)}{N}$: sample relative frequency of ω_i.

If the articles are used randomly throughout the text, their sample relative frequencies should be approximately the same for any sample size N, i.e., their sample relative frequencies should show up as a horizontal line in a graph of $p(i, N)$ as a function of N. Figure 1.2, however, reveals that the observed sample relative frequencies do not show up as horizontal lines. Instead, they reveal non-random developmental profiles. The indefinite article *a* (panel A) is used more intensively in the central sections of the book than in the beginning or end. The definite article *the* (panel B) shows a more or less linear increase in relative sample frequency.

These developmental profiles might be due to chance. What is the probability that a random re-ordering of the words of *Alice in Wonderland* would lead to a similar pattern? We can approach this question by calculating the mean sample relative frequencies at twenty equally-spaced intervals for a large number of random permutations of the order of the words in *Alice in Wonderland*. As we shall see in more detail in Chapter 5, in a random permutation of the text the effects of cohesion in word use at the levels of sentence and discourse is eliminated. If the values actually observed for the text are more extreme than those observed for 95% of the permutation runs, then we know that the probability that the observed pattern arose by chance is less then 0.05.

Panels A and B of Figure 1.2 show the result of this randomization test for a total of 5000 permutations runs. For both articles, the average proportion or Monte Carlo mean, represented by a solid line, is constant, exactly what we expect when the tokens of a word are equally spread out over the text. The dotted lines mark the two-tailed 95% Monte Carlo confidence interval. Panel A shows that for measurement points 8–10, 13–14, and 16–19 the observed values for $p(a, N)$ fall outside this confidence interval. Apparently, *a* tends to be slightly overrepresented in the second half of the text. Turning to panel B, we find that with the exception of the final measurement point (the full text), all observed relative sample frequencies of *the* are well below the lower 95% confidence limit. Again we may conclude that the observed pattern for *the* is far from random.

The empirical developmental profiles of the articles are possibly linked with narrative development in *Alice in Wonderland* in the following way. In the initial sections of the book, new participants and new scenes are introduced, but as one continuous reading, Alice re-encounters participants and revisits places where she had been before. Since the indefinite article typically introduces new information and the definite article given information, the increase in the use of *the* and the decrease in the use of *a* in the second half of the text might be the consequence of thematic development in the narrative. Note, however, that this leaves the increase in the relative frequency of *a* in the first half of the text unexplained.

Summing up, in statistical analyses of textual data it is important to realize that the values of simple statistics such as means and proportions are heavily influenced by the sample size, for two reasons. First, the law of large numbers cannot be relied on when dealing with words and their frequencies of use. Second, authors do not use words at random. Word usage reflects lexical

cohesion both at the level of the sentence and at the level of discourse. These two factors should always be kept in mind when comparing the quantitative properties of textual materials.

1.2 The frequency spectrum

We have seen that statistics such as the sample mean and median increase when the sample size is increased. The variability of the sample mean has severe consequences for the comparison of texts. As illustrated for *Alice in Wonderland* and *Through the looking-glass*, it is possible to adjust for differences in sample size by considering a subset of the word tokens in the larger set. Such a procedure, however, has some obvious drawbacks. First, data in the larger sample have to be discarded, which implies a loss of information. Second, there are no good criteria available for deciding which tokens in the larger text should be discarded. Especially for novels and cohesive texts in general, the removal of any part of the text is completely arbitrary. Not surprisingly, considerable effort has been spent on the development of quantitative measures that characterize textual properties independently of sample size. This chapter reviews a series of such statistics. Unfortunately, the main thrust of the argument is a negative one: Almost all 'constants' proposed in the literature reveal specific developmental profiles in sampling time just as the sample mean and median. Before discussing a number of measures that have been put forward as text constants, I should first introduce the concept of the **grouped frequency distribution** or **frequency spectrum**.

Word frequencies in *Alice in Wonderland* range from 1 to 1631. The most frequent word is *the*, and this is the only word with this particular token frequency. Conversely, the lowest token frequency, 1, is represented by 1176 different words. The words which occur once only in a text are known as **hapax legomena**, from Greek *hapax*, 'once', and *legomenon*, 'read'. Typically, 1 is the frequency that is represented by the greatest number of words. The number of words that occur twice in *Alice in Wonderland*, 402, the so-called **dis legomena**, is substantially smaller, but in its turn almost twice the number of words that occur three times, 233. I will use the index m to denote these frequency classes. The number of word types in a given frequency class for a sample of size N will be denoted by $V(m,N)$. Thus, $V(1,N)$ denotes the number of hapax legomena, $V(2,N)$ the number of dis legomena, etc.

Definition 1.5 m: index for frequency class.

Definition 1.6 $V(m,N) = \sum_{i=1}^{V(N)} \mathrm{I}_{[f(i,N)=m]}$: the number of types with frequency m in a sample of N tokens.

The identity operator $\mathrm{I}_{[\alpha]}$ that appears in the definition of $V(m,N)$ yields the value 1 if the expression α is true, and zero otherwise. Note that N and $V(N)$ can be expressed in terms of m and $V(m,N)$:

$$N = \sum_m mV(m,N) \tag{1.1}$$

Table 1.3: *The frequency spectrum* $V(m,N)$ *of* Alice in Wonderland.

m	$V(m,N)$	m	$V(m,N)$	m	$V(m,N)$	m	$V(m,N)$
1	1176	31	3	62	1	144	1
2	402	32	4	63	1	145	1
3	233	33	4	67	2	148	1
4	154	34	3	68	4	151	1
5	99	35	4	73	1	153	1
6	57	37	1	74	1	170	1
7	65	38	4	75	1	177	1
8	52	39	4	77	2	179	1
9	32	40	4	79	1	182	1
10	36	41	2	80	1	194	1
11	23	42	2	81	1	211	1
12	20	43	2	82	2	247	1
13	34	44	1	83	2	263	1
14	20	45	4	85	1	280	1
15	12	46	1	87	1	356	1
16	9	47	1	88	2	364	1
17	9	48	1	90	1	365	1
18	10	49	4	93	1	386	1
19	8	50	2	94	1	410	1
20	5	51	4	96	2	460	1
21	6	52	3	98	1	510	1
22	3	53	1	102	2	528	1
23	3	54	3	108	1	540	1
24	6	55	3	113	1	629	1
25	9	56	1	114	1	726	1
26	4	57	2	121	1	866	1
27	6	58	2	128	1	1631	1
28	3	59	1	131	1		
29	6	60	2	133	1		
30	6	61	3	136	1		

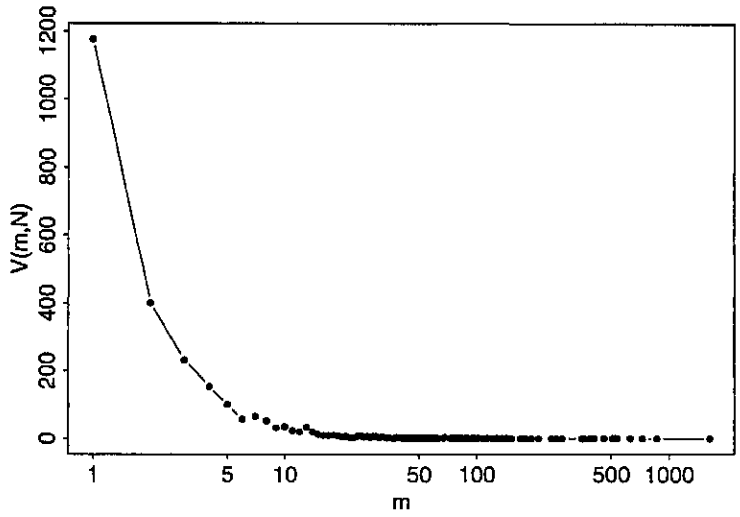

Figure 1.3: *The frequency spectrum of* Alice in Wonderland *(m: frequency class; $V(m,N)$: number of types with frequency m).*

$$V(N) = \sum_m V(m,N). \tag{1.2}$$

Table 1.3 lists $V(m,N)$ for the $N = 26505$ tokens of *Alice in Wonderland*, and Figure 1.3 visualizes this grouped frequency distribution or **frequency spectrum**. The horizontal axis plots the frequency classes m using a logarithmic scale. The vertical axis plots $V(m,N)$. Note that $V(m,N)$ is a rapidly decreasing function of m with a long tail of high frequencies m that are instantiated by very few types.

The curve of the frequency spectrum is smooth enough to suggest that in theory $V(m,N)$ might be a monotonically decreasing function of m for which the inequality

$$V(m,N) > V(m+1,N) \tag{1.3}$$

holds. The irregularities observed from $m = 6$ onwards (see also Table 1.3) would then be due to sampling error. Highly skewed frequency spectra of this kind are typical in lexical statistics.

Closely related to the grouped frequency distribution $V(m,N)$ is the so-called **empirical structural type distribution** $g(m,N)$, which specifies the number of different word types which occur m or more times in a sample of N tokens.

Definition 1.7 $g(m,N) = \sum_{i=1}^{V(N)} \mathrm{I}_{[f(i,N) \geq m]}$: the number of types with a frequency m or more in a sample of N tokens.

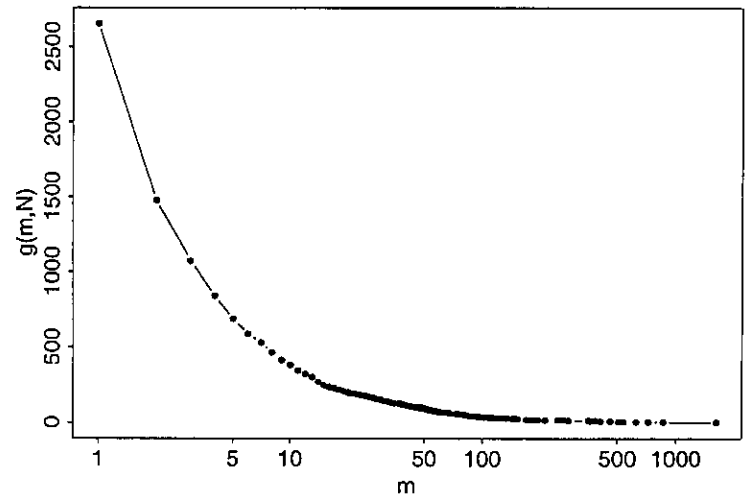

Figure 1.4: *The empirical structural type distribution of* Alice in Wonderland *(m: frequency class; $g(m,N)$: number of types occurring m or more times).*

Table 1.4 presents the empirical structural type distribution of *Alice in Wonderland*, and Figure 1.4 plots the corresponding empirical structural type distribution, again using a logarithmic scale for the horizontal axis. Note that $g(1,N)$ is equal to $V(N)$, and that for *the*, the highest-frequency word in the text, $g(1631,N) = 1$.

The grouped frequency distribution and the empirical structural type distribution are related by the following expressions:

$$V(m,N) = g(m,N) - g(m+1,N), \tag{1.4}$$

$$g(m,N) = \sum_{w \geq m} V(w,N). \tag{1.5}$$

Table 1.4: *The empirical structural type distribution* $g(m, N)$ *of* Alice in Wonderland.

m	$g(m,N)$	m	$g(m,N)$	m	$g(m,N)$	m	$g(m,N)$
1	2651	31	143	62	67	144	27
2	1475	32	140	63	66	145	26
3	1073	33	136	67	65	148	25
4	840	34	132	68	63	151	24
5	686	35	129	73	59	153	23
6	587	37	125	74	58	170	22
7	530	38	124	75	57	177	21
8	465	39	120	77	56	179	20
9	413	40	116	79	54	182	19
10	381	41	112	80	53	194	18
11	345	42	110	81	52	211	17
12	322	43	108	82	51	247	16
13	302	44	106	83	49	263	15
14	268	45	105	85	47	280	14
15	248	46	101	87	46	356	13
16	236	47	100	88	45	364	12
17	227	48	99	90	43	365	11
18	218	49	98	93	42	386	10
19	208	50	94	94	41	410	9
20	200	51	92	96	40	460	8
21	195	52	88	98	38	510	7
22	189	53	85	102	37	528	6
23	186	54	84	108	35	540	5
24	183	55	81	113	34	629	4
25	177	56	78	114	33	726	3
26	168	57	77	121	32	866	2
27	164	58	75	128	31	1631	1
28	158	59	73	131	30		
29	155	60	72	133	29		
30	149	61	70	136	28		

1.3 Zipf

Among the earliest studies on word frequency distributions the work by Zipf (1935, 1949) figures prominently. Zipf ordered the words in his texts by decreasing frequency, and considered the relation between rank order and frequency. Consider Table 1.5, which lists the twenty most frequent words in

Table 1.5: *The twenty most frequent words in* Alice in Wonderland *ordered according to decreasing frequency.*

z	$f_z(z,N)$	word	z	$f_z(z,N)$	word
1	1631	*the*	11	365	*in*
2	866	*and*	12	364	*you*
3	726	*to*	13	356	*was*
4	629	*a*	14	280	*that*
5	540	*she*	15	263	*as*
6	528	*it*	16	247	*her*
7	510	*of*	17	211	*at*
8	460	*said*	18	194	*on*
9	410	*I*	19	182	*all*
10	386	*Alice*	20	179	*with*

Alice in Wonderland in their Zipfian rank order. The most frequent word, *the*, is assigned the **Zipf rank** $z = 1$, the next most frequent word, *and*, is assigned rank $z = 2$, and so on. Words with the same frequency are arranged in some arbitrary order and they receive successively larger Zipf ranks. For instance, the 1176 hapax legomena in *Alice in Wonderland* are assigned the Zipf ranks 1476, 1477, 1478, ..., 2651. (This implies that the actual Zipf rank of a hapax legomenon is not of interest, but rather the ranks at which the first and last hapax legomenon are observed.) I will use the notation $f_z(z, N)$ for the frequency of a word with Zipf rank z. Thus $f_z(1, N)$ is the frequency of the word with Zipf rank 1, the subscript indicating that the frequency is to be understood as with respect to a Zipfian ranking.

Definition 1.8 z: Zipf rank in a word list ordered by decreasing frequency.

Definition 1.9 $f_z(z, N)$: frequency in a sample of N tokens of a word with Zipf rank z.

The Zipfian **rank-frequency distribution** is the inverse of the empirical structural type distribution:

$$g(m, N) = z \Leftrightarrow f_z(z, N) = m\,. \tag{1.6}$$

For instance, for the highest-frequency word in *Alice in Wonderland*, *the*, we have

$$g(1631, N) = 1,$$

but at the same time

$$f_z(1, N) = 1631.$$

Similarly, at the low-frequency end of the frequency spectrum we have:

$$\begin{aligned} g(1, N) &= 2651, \\ f_z(2651, N) &= 1. \end{aligned}$$

In general, if a word with frequency m has Zipf rank z, then the frequency ordering underlying the Zipf ranking implies that there are z words with at least frequency m, which in turn implies that $g(m, N) = z$.

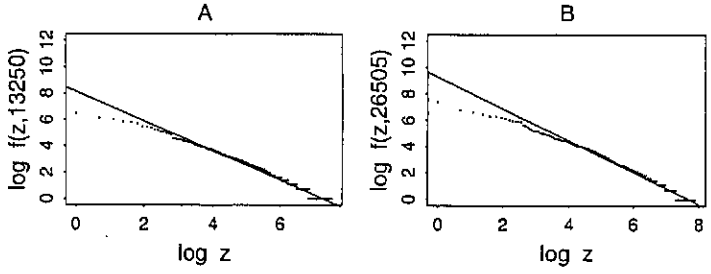

Figure 1.5: *Word frequency $f_z(z, N)$ as a function of Zipf rank z in the double logarithmic plane for $N = 13250$ (panel A) and for $N = 26505$ (panel B).*

Panel A of Figure 1.5 is a points plot of $\log f_z(z, N)$ against $\log z$ for the first 13250 tokens of *Alice in Wonderland*. The solid line is the corresponding least squares regression line. Note that the highest frequencies appear as individual points at the left hand edge of the plot, and that the large numbers of hapax legomena and dis legomena appear as horizontal line segments at the right-hand edge of the plot. The corresponding plot in panel B for the complete text reveals a highly similar pattern. Zipf observed such roughly linear plots for many different kinds of texts. This led him to formulate the following relation between $f_z(z, N)$ and z:

$$f_z(z, N) = \frac{C}{z^a}. \tag{1.7}$$

In (1.7), a is the free parameter of the model that determines the slope of the regression lines in Figure 1.5, C is a normalizing constant. Taking logarithms at both sides, the linear relation between $\log f_z(z, N)$ and $\log z$ follows immediately:

$$\begin{aligned} \log f_z(z, N) &= \log \frac{C}{z^a} \\ &= \log C - a \log z. \end{aligned}$$

This inverse relation between log Zipf rank and log frequency is known as **Zipf's law.**

Thus far, we have considered Zipf's law in terms of the absolute sample frequencies of words and their Zipf ranks. Absolute sample frequencies, however, are subject to sampling error, and will therefore diverge slightly from Zipf's law. Theoretically, the frequency of a word is expected to be $N\pi_i$, with π_i the population probability of ω_i (see section 2.2). Underlying the observed frequencies, there is a distribution of probabilities for which Zipf's law should also be valid. Let's therefore reformulate Zipf's law in terms of the probabilities of words. Instead of focusing on $f_z(z, N)$, we first consider the corresponding **relative sample frequencies** $f_z(z, N)/N$. Assume furthermore, for the sake of the argument, that these sample relative frequencies are reliable estimates of the population probabilities.

Definition 1.10 π_z: probability of the word ω_z with Zipf rank z.

Definition 1.11 $\hat{\pi}_z$: the probability of word ω_z estimated from its sample relative frequency: $\hat{\pi}_z = f_z(z, N)/N$.

We can now reformulate Zipf's law as

$$\pi_z = \frac{C}{z^a}. \tag{1.8}$$

In both (1.7) and (1.8), C is a normalizing constant. Its function in (1.8) is to ensure that the probabilities sum up to unity:

$$\sum_z \pi_z = 1.$$

In (1.7), it ensures that the frequencies $f_z(z, N)$ sum up to N:

$$\sum_z f_z(z, N) = N.$$

(Note that this implies that the value of the normalizing constant in (1.7) is N times that in (1.8). Going from frequencies to estimated probabilities therefore implies a leftward shift by $\log N$ in the double logarithmic plane.)

The probability distribution (1.8) is known as the **zeta distribution,** which owes its name to the Riemann Zeta function (after the German mathematician G.F.B. Riemann)

$$\zeta(a) = 1 + \left(\frac{1}{2}\right)^a + \left(\frac{1}{3}\right)^a + \ldots + \left(\frac{1}{n}\right)^a + \ldots .$$

Apart from the (normalizing) constant C, the successive terms of the Riemann Zeta function spell out the probabilities of the words with Zipf rank $z = 1, 2, \ldots$. In other words, the probability of a word with the n-th rank is given by the n-th term in the expansion of $\zeta(a)$. Thus we can restate the zeta

function in terms of the Zipf probabilities (1.8):

$$\begin{aligned}\zeta(a) &= 1+\left(\frac{1}{2}\right)^a+\left(\frac{1}{3}\right)^a+\ldots+\left(\frac{1}{n}\right)^a+\ldots \\ &= (\pi_1+\pi_2+\pi_3+\ldots+\pi_n+\ldots)\frac{1}{C}.\end{aligned}$$

Zipf often took a to equal unity, in which case the zeta function reduces to the so-called harmonic series

$$\sum_z \frac{1}{z} \quad (z=1,2,3,\ldots).$$

Zipf (1935) refers to the corresponding probability distribution

$$\pi_z = \frac{C}{z}$$

as the **standard harmonic distribution.**[2]

Given the standard harmonic distribution, $V(m,N)$ can be expressed as a function of m:

$$V(m,N) \propto \frac{1}{m(m+1)}. \tag{1.9}$$

To see this, consider two Zipf ranks z_1 and z_2 such that $f_z(z_1,N) = m+1$ and $f_z(z_2,N) = m$, with $m > 0$, where we choose z_1 such that there is no $z > z_1$ for which $f_z(z,N) = m+1$, and similarly z_2 such that there is no $z > z_2$ for which $f_z(z,N) = m$. In other words, we focus on the jumps in the rank-frequency step function, as illustrated graphically in Figure 1.6, and numerically for $m = 1,2,3$ in *Alice in Wonderland* in Table 1.6. Crucially, we can write

$$V(m,N) = z_2 - z_1.$$

Since $f_z(z_2,N) = \frac{C}{z_2} = m$ we have $z_2 = \frac{C}{m}$. Likewise, $z_1 = \frac{C}{m+1}$, and hence

$$V(m,N) = z_2 - z_1 = \frac{C}{m} - \frac{C}{m+1} = \frac{C}{m(m+1)}.$$

[2]The harmonic distribution does not converge. Because the probabilities π_z do not sum up to unity, it is not a proper probability distribution. However, since

$$\sum_{i=1}^{V} \frac{1}{i} - \log(V) \cong \gamma$$

for $V \to \infty$, with $\gamma = 0.57723$ (Euler's constant), we have that

$$\frac{1}{\log(V)+\gamma}\sum_{i=1}^{V}\frac{1}{i} = 1,$$

which allows us to use the harmonic distribution as an approximation for a probability distribution for sufficiently large V.

Since the number of hapax legomena tends to be roughly half the vocabulary size, the normalizing constant C is often taken to be $V(N)$:

$$V(m,N) = \frac{V(N)}{m(m+1)}. \tag{1.10}$$

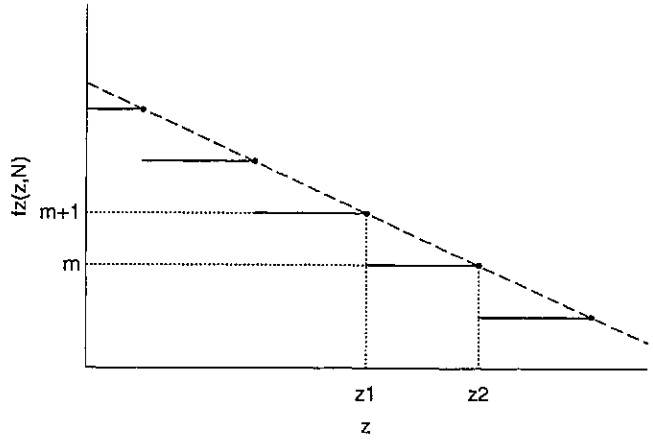

Figure 1.6: *The rank-frequency distribution as a step function in the double logarithmic plane and the relation with the elements of the frequency spectrum:* $V(m,N) = z_2 - z_1$.

Zipf (1935:47) hoped that the standard harmonic distribution would provide

> ...Dynamic Philology with a *standard curve of distribution* in reference to which the frequency distribution of any other language can be described. If the curve of the frequency distribution of a given language conforms at any point with the standard harmonic curve or if it deviates at any point either slightly or seriously above or below, these facts may shed welcome light on significant factors in the structure of language.

Zipf's hopes have been partially fulfilled. It is clear that in panel A of Figure 1.5 the frequencies of the lowest ranks deviate substantially from their theoretical values according to (1.7). Subsequent research by Mandelbrot (1953) has suggested that this deviation can be captured by introducing a second free parameter in the model, a proposal to which we discuss in more detail in section 3.2.3. Finally, the words occurring with the highest Zipf ranks are typically function words such as *the, a, for,* and *she* that have properties that differ fundamentally from content words such as *sister, white,* and *rabbit.*

The usefulness of Zipf's model, however, is severely limited because its parameters are highly dependent on the sample size N. To see this, consider again the graphs shown in Figure 1.5. Panel B plots the Zipfian curve

Table 1.6: *The relation between the Zipf rank, the empirical structural type distribution, and the spectrum elements, illustrated for* $m = 1, 2, 3$ *in* Alice in Wonderland *(*$N = 26505$*).*

Zipf rank		Number of spectrum elements
841		
⋮		233 tris legomena $= g(3,N) - g(4,N)$
1073	$= g(3,N)$	
1074		
⋮		402 dis legomena $= g(2,N) - g(3,N)$
1475	$= g(2,N)$	
1476		
⋮		1176 hapax legomena $= g(1,N) - g(2,N)$
2651	$= g(1,N)$	

for N =26505, twice the sample size of panel A. The general shapes of the two curves are highly similar, although the divergence from linearity at the left hand side seems to have increased somewhat for the full text (panel B). More disconcerting is the observation that the slope increases from -1.119 for N =13250 to -1.205 for N =26505, and that the changes in the parameters of the model as a function of N are systematic, as shown in Figure 1.7. Panel A plots the intercept as a function of the sample size N, and panel B the slope. The intercept is an increasing function of N, the slope is a decreasing function of the sample size. Clearly, the parameters of the zeta distribution are subject to the same kind of systematic variation as the sample mean frequency. A way to take this variability into account in a principled way will be presented in Chapter 3.

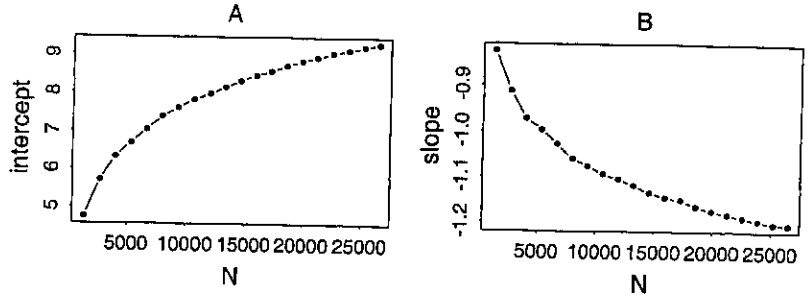

Figure 1.7: *The dependency of the two parameters of Zipf's zeta distribution, the intercept (panel A) and the slope (panel B) on the sample size* N*, plotted at 20 equally-spaced intervals for* Alice in Wonderland *(compare 1.5).*

Thus far, we have considered Zipf's law as formulated for the rank-frequency distribution. Zipf also proposed to analyze the frequency spectrum itself in terms of the zeta distribution. When we plot $V(m,N)$ against m in the double logarithmic plane, as shown for *Alice in Wonderland* in the left panel of Figure 1.8, we again find a roughly linear relation. Zipf (1935:40–44) points out that a model of the form

$$V(m,N) \propto \frac{1}{m^a} \tag{1.11}$$

is accurate primarily for the smaller values of m. This point is highlighted by the dashed line, which represents a linear fit $\log(V(m,N)) = a + b\log(m)$ based on the first 15 ranks. For these first 15 ranks, the fit seems quite reasonable, but it clearly does not capture the pattern among the higher-frequency ranks, which tend to have higher values than expected. When we base a linear fit on the full spectrum, the regression line, represented by a dotted line, deviates considerably from the observed lowest-frequency ranks.

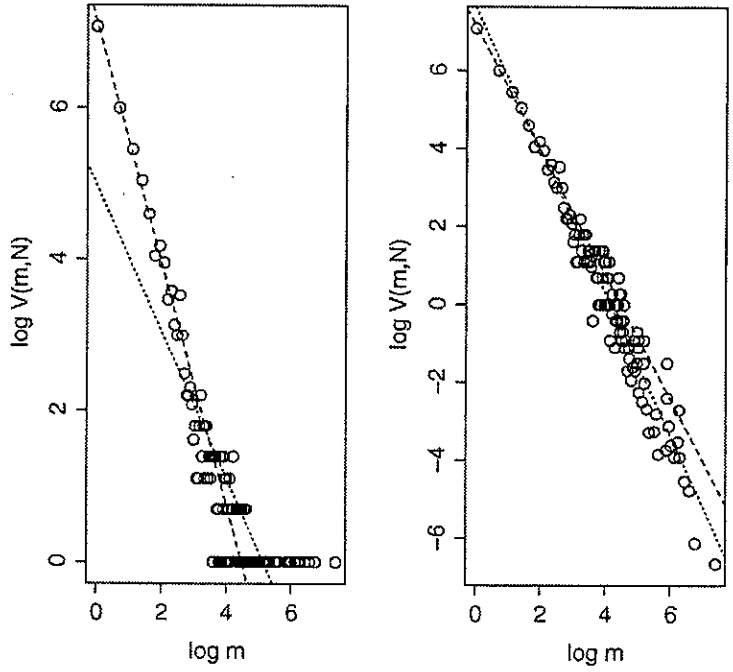

Figure 1.8: *The frequency spectrum of* Alice's Adventures in Wonderland *in the double logarithmic plane. Left panel: the integer-valued spectrum with linear fits for* $m = 1 \ldots 15$ *(dashed line) and for the complete range of* m *(dotted line). Right panel: the same spectrum elements after transformation into fractional values for higher-frequency ranks, with linear fits for* $m = 1 \ldots 15$ *(dashed line) and for the complete range of* m *(dotted line).*

In part, the problem that we are dealing with is a discretization problem. Word frequencies are integers, yet the power model

$$V(m,N) = am^b$$

which yields the straight line $\log(V(m,N)) = \log(a) + b\log(m)$ in the bi-logarithmic plane, expects words to have real-valued frequencies. Instead of having sparsely populated high-frequency ranks m with words occuring with integer frequencies, the model assumes that all ranks are populated with fractional numbers of types that become smaller as m increases. By itself, this need not be a problem, as long as there is a way to transform an integer-valued distribution into a real-valued distribution.

Church and Gale (1991) and Gale and Sampson (1995) propose the following technique to obtain fractional spectrum elements for the sparsely populated higher frequency ranks m. Let m_p and m_f denote the ranks for which $V(m,N) > 0$ that immediately precede and follow rank m. If $V(m-1,N) > 0$, we have that $m_p = m-1$, otherwise, $m_p < m-1$. Likewise, if $V(m+1,N) > 0$, then $m_f = m+1$, otherwise, it will be greater than $m+1$ due to intervening zero ranks. We can now define the real-valued $V_r(m,N)$ as follows:

$$V_r(m,N) = \begin{cases} V(1,N) & \text{if} \quad m = 1 \\ \frac{2V(m,N)}{m_f - m_p} & \text{if} \quad 1 < m < \max(m), \\ \frac{2V(m,N)}{2m - m_p} & \text{if} \quad m = \max(m). \end{cases} \tag{1.12}$$

When zero ranks intervene between m and m_p or m_f, the difference $m_f - m_p$ will be greater than 2, so that $V_r(m,N) < V(m,N)$, otherwise $V_r(m,N)$ will be the same as $V(m,N)$ The right-hand panel of Figure 1.8 plots the resulting $V_r(m,N)$ for *Alice's Adventures in Wonderland* in the bi-logarithmic plane using circles, which now approximately follow a straight line. The downward curvature in the observed values for the lowest frequency ranks is not eliminated, however, as shown by the dashed line, a least squares regression line to the first 15 ranks. In contrast to the dotted line in the left panel, the dotted line in the right panel, the least-squares regression line using all observations, seems reasonable for all but the first two and last two ranks.

Definition (1.12) of $V_r(m,N)$ sets $V_r(m,N)$ to zero whenever $V(m,N) = 0$. Thus, following Church and Gale (1991), the right-hand panel of Figure 1.8 plots exactly the same number of points as the left panel of Figure 1.8. But because some integer-valued spectrum elements have been replaced by smaller real-valued spectrum elements, we have that

$$\sum_m V_r(m,N) < \sum_m V(m,N) = V,$$
$$\sum_m mV_r(m,N) < \sum_m mV(m,N) = N.$$

The discrepancy for $V(N)$, however, is easily solved by defining $V_r(m,N)$ for zero ranks as well. Let m_0 denote a rank for which $V(m,N) = 0$, and let ${}_pm$ denote the greatest nonzero rank smaller than m_0 and ${}_fm$ the smallest nonzero rank greater than m_0. Then

$$V_r(m_0,N) = \frac{V_r(m_p,N) + V_r(m_f,N)}{m_f - m_p}, \tag{1.13}$$

the average of the nonzero spectrum elements surrounding m_0. With the addition of $V_r(m_0,N)$, the discrepancy between $V(N)$ and $\sum_m V_r(m,N)$ is removed. To see why this is so, consider a spectrum element m with $k = m_f - m_p - 2$ surrounding zero spectrum elements. The $V(m,N)$ types of rank m are reset to $2V(m,N)/(k+2)$, leaving $kV(m,N)/(k+2)$ fractional types which we divide equally among each of the k empty ranks. In this way, all $V(m,N)$ types are still present in the distribution, but now divided over $k+1$ ranks instead of being concentrated in one rank m only. An empty rank m_0 therefore receives $V(m_p,N)/(k+2)$ fractional types from its nearest left nonzero rank, and likewise $V(m_f,N)/(k+2)$ fractional types from its nearest right nonzero rank, which immediately leads to (1.13). There is no guarantee, however, that $\sum_m mV_r(m,N)$ will equal N. For our present example of *Alice's Adventures in Wonderland*, $\sum_m mV_r(m,N) = 26893.61$ instead of 26505, even though $\sum_m V_r(m,N)$ is now identical to $V(N)$. This implies that the values of $V_r(m,N)$ are slightly too high.

Figure 1.8 illustrates that fitting a straight line to the real-valued approximate spectrum elements may not do justice to the slight downward curvature at the head of the spectrum. The left panel of Figure 1.9 shows that this curvature is handled in a more principled way when we smooth the spectrum using (1.10), with $V(26505) = 2651$:

$$V_z(m,N) = \frac{2651}{m(m+1)}.$$

Evaluating the goodness of fit in terms of the mean squared error (MSE) for the first 40 ranks,

$$\text{MSE}_{(40)} = \frac{\sum_{i=1}^{40}(V(m_i,N) - V_z(m_i,N))^2}{40},$$

we find that using (1.10) instead of a simple linear fit reduces the MSE from 30641.48 to 648.23. Although qualitatively and quantitatively a substantial improvement, the MSE remains quite high, and the left panel of Figure 1.9 shows that the curve of $V_z(m,N)$ tends to be too low for the higher ranks ($m > 20$).

A more flexible Zipfian smoother has been proposed by Naranan and Balasubrahmanyan (1998),

$$V_{nb}(m,N) = \frac{Ce^{-\mu/m}}{m^\gamma}, \tag{1.14}$$

which shares the term $1/m^\gamma$ with the simple Zipfian power function (1.11), but adds the term $e^{-\mu/m}$ to handle the downward curvature at the head of the frequency spectrum. The right panel of Figure 1.9 plots the fit of this model

to the data. Evaluated in terms of the MSE for the first 40 ranks, 77.37, we observe a substantial improvement in goodness of fit. (For details on how the parameters of (1.14) can be estimated, see section 3.4 in Chapter 3.)

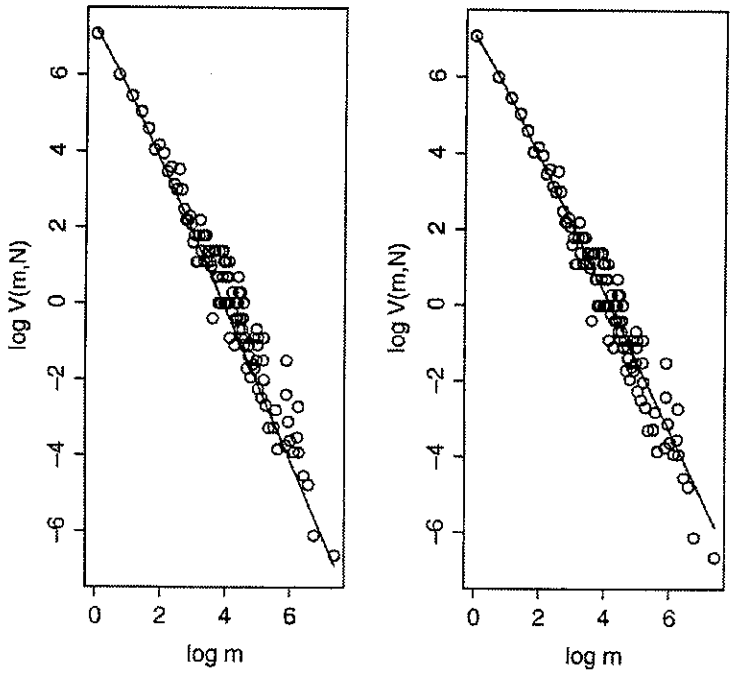

Figure 1.9: *The real-valued approximate frequency spectrum of* Alice's Adventures in Wonderland *in the double logarithmic plane with a simple Zipfian fit (left panel) and a Naranan-Balasubrahmanyan Zipfian fit (right panel).*

Although the Naranan-Balasubrahmanyan Zipfian model generally provides very good fits to empirical spectra, it suffers from the same problem as observed in Figure 1.7 for the parameters of Zipf's rank-frequency distribution, namely, systematic changes as a function of the sample size N. Figure 1.10 illustrates this systematic variability for a Dutch text, *Max Havelaar*, by Multatuli, the pseudonym of Eduard Douwes Dekker (1820-1887). The words of this text were re-arranged in a random order to eliminate possible effects of non-randomness in word use (see Chapter 5 for detailed discussion of the randomness assumption). The upper left panel plots $V(N)$ (upper circles), $V(1,N)$ (central circles), and $V(2,N)$ (bottom circles) as a function of N at 20 equally-spaced intervals. The remaining panels plot C (upper right), μ (lower left) and γ (lower right) as a function of N. Throughout the text, C appears to increase with N. During the first 8 measurement points, μ increases, after which it becomes more or less stable. From measurement point 9, γ emerges as a decreasing function of N. Thus, at least two parameters have to be adjusted to accommodate a change in sample size.

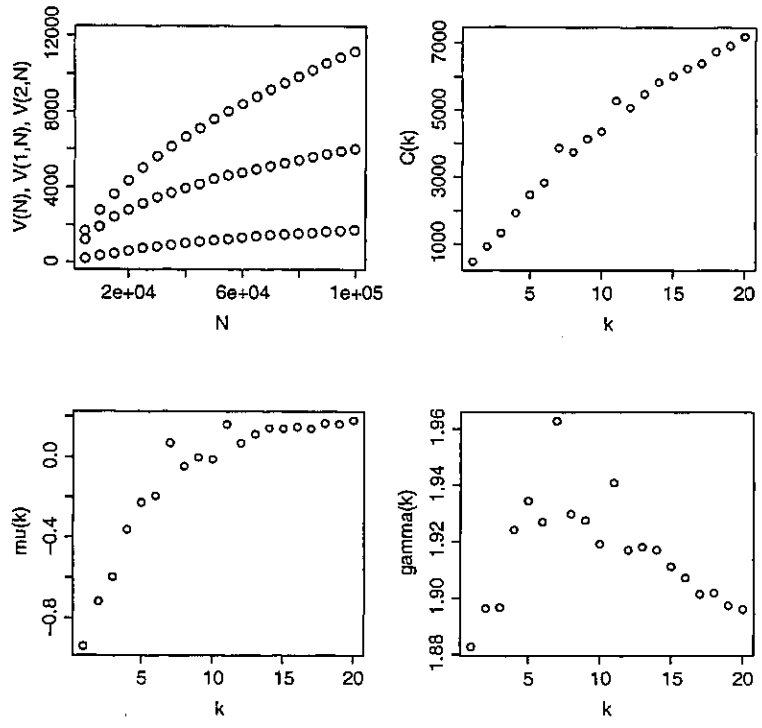

Figure 1.10: *The dependency of the three parameters of the Naranan-Balasubrahmanyan Zipfian model as a function of the sample size N in* Max Havelaar *by Multatuli. The upper right panel plots C as a function of N, the lower left panel μ as a function of N, and the lower right panel γ as a function of N. The upper left panel plots $V(N)$ (upper circles), $V(1,N)$ (central circles), and $V(2,N)$ (lower circles) as a function of N at 20 equally-spaced intervals.*

To see why the parameters of Naranan-Balasubrahmanyan Zipfian model change systematically as a function of N, consider the ratio of any two spectrum elements,

$$R(m,n) = V(m,N)/V(n,N).$$

Given Naranan and Balasubrahmanyan's model, we have that

$$\begin{aligned} R_{nb}(m,n) &= \frac{Ce^{-\mu/m}m^{-\gamma}}{Ce^{-\mu/n}n^{-\gamma}} \\ &= e^{-\frac{\mu(n-m)}{mn}}\left(\frac{m}{n}\right)^{-\gamma}. \end{aligned}$$

Note that C disappears in R_{nb}, which leaves us with two parameters to account for the kind of changes in the empirical values of these ratios illustrated

for *Alice's Adventures in Wonderland* and *Max Havelaar* in Figure 1.11. The circles represent $R(2,1)$, the triangles $R(3,2)$, and the solid lines the corresponding expected values (using binomial interpolation, see section 2.6). Because the empirical ratios change systematically with N, the parameters μ and γ have to be adjusted continuously.

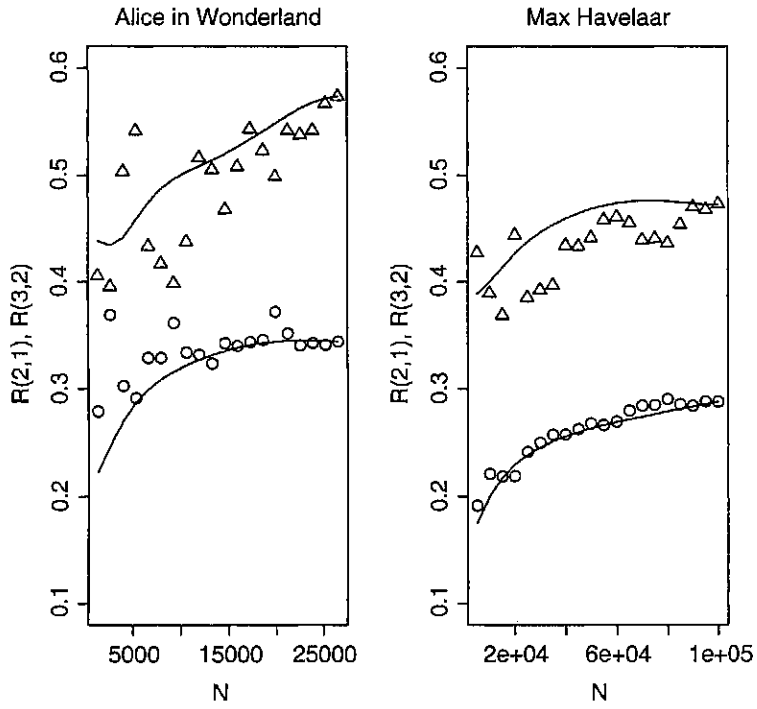

Figure 1.11: *The spectrum ratios $V2/V1$ ($R(2,1)$, circles) and $V3/V2$ ($R(3,2)$, triangles) for* Alice's Adventures in Wonderland *left panel) and* Max Havelaar *(right panel) for 20 equally spaced sample sizes N. The solid lines represent the corresponding expected values.*

1.4 The quest for characteristic constants

We have seen that lexical measures such as mean frequency and vocabulary size, as well as the parameters of the zeta distribution and the Naranan-Balasubrahmanyan Zipfian model all depend on the sample size. This state of affairs leads to severe problems when texts of different lengths have to be compared. Not surprisingly, a great many measures have been proposed as text constants, measures that do not vary with the size of the sample.

The oldest of these measures was developed by Yule (1944), in his seminal book, *The Statistical Study of Literary Vocabulary*. In this comprehensive study of the frequency distributions of a great number of different texts, all compiled by hand on individual slips of paper, Yule proposed a quantitative textual measure that, apart from sampling fluctuations, should be independent of sample size. His so-called characteristic constant K,

$$K = 10000 \frac{\sum_m m^2 V(m,N) - N}{N^2}, \tag{1.15}$$

is a measure of the rate at which words are repeated in a text. To see this, it is convenient to write K as follows:

$$K = 10000 \left\{ \sum_m [V(m,N) \frac{m}{N} \frac{m}{N}] - \frac{1}{N} \right\}.$$

The factor 10000 is a scale factor, introduced only to avoid overly small values of K that might otherwise be difficult to read. The proportion m/N is the sample estimate of the probability of sampling a word with token frequency m. Hence, m^2/N^2 is the probability of sampling such a word twice in a row, assuming that the word probabilities are constant (sampling with replacement).

A closely related measure has been proposed by Simpson (1949):

$$D = \sum_m V(m,N) \frac{m}{N} \cdot \frac{m-1}{N-1}. \tag{1.16}$$

Consider a word ω_i with frequency m in a sampling situation without replacement. The probability that the very first word sampled is precisely ω_i equals m/N. The probability that the next token sampled represents this same type is given by $(m-1)/(N-1)$: the number of remaining tokens of ω_i divided by the total number of remaining tokens. Thus $\frac{m}{N}\frac{m-1}{N-1}$ estimates the likelihood that two tokens of the same type are sampled in succession. The value of D is obtained by summation of this likelihood for all $V(N)$ types.

Both K and D are measures of the repeat rate. In Chapter 2, we shall see that they can also be viewed as weighted average probabilities. Both are heavily influenced by the highest frequency words, for which the probability of repeated use is greatest. The large dots in panels A and B of Figure 1.12 show that in *Alice in Wonderland* K and D reveal highly similar developmental patterns. Apparently, the repeat rate in this text is high in the initial sections, decreases first, and then slowly increases again.

Doubts concerning the stability of K and D for varying sample sizes have led to the formulation of other text constants. Recall that the mean frequency $N/V(N)$ varies with the sample size N. Would it be possible to eliminate this variation by considering simple functions of N and $V(N)$? Guiraud (1954) proposed a measure in which the square root of the sample size replaces the sample size in what is known as the type-token ratio, $V(N)/N$, the inverse of the mean type frequency, as follows:

$$R = \frac{V(N)}{\sqrt{N}}, \tag{1.17}$$

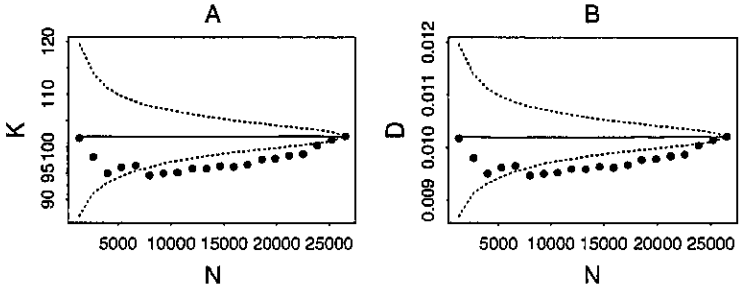

Figure 1.12: *The characteristic constants K (panel A) and D (panel B) as a function of the sample size N in* Alice in Wonderland. *The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.*

and Brunet (1978) suggested a power relation between N and $V(N)$,

$$W = N^{V(N)^{-a}}. \tag{1.18}$$

If R and W are truly constants, then $V(N)$ reduces to very simple functions of N:

$$\begin{aligned} V(N) &= R\sqrt{N}\,, \\ V(N) &= \left(\frac{\log(W)}{\log(N)}\right)^{-\frac{1}{a}} = (\log_N W)^{-\frac{1}{a}}\,. \end{aligned}$$

But are R and W truly constant and independent of N? The large dots in Figure 1.13, which plot the observed values of R and W for 20 equally-spaced values of N in *Alice in Wonderland*, show that this is not the case. In this novel, R and W are increasing functions of the sample size N. Unlike K and D, R and W have no probabilistic interpretation. The value of the parameter a in the expression for W, for instance, has no sensible interpretation and is usually fixed at 0.17, a heuristic value that has been found to produce the desired result of producing a roughly constant relation between N and $V(N)$.

It is easy to see why R and W change with N when we consider a population with a fixed number of types, for instance, a distribution of 10000 tokens sampled from a population with 50 equiprobable types, as shown in Figure 1.14. Each panel displays 40 measurement points that are 250 tokens apart. The upper left panel shows that all 50 types already appear in the sample once 250 tokens have been sampled. The upper right panel shows the linear increase of the average token frequency in the interval $N = [250, 10000]$. The lower left panel shows that R decreases as N is increased. Because V is fixed, the plot effectively shows the function $y = 50x^{-1/2}$. The lower right

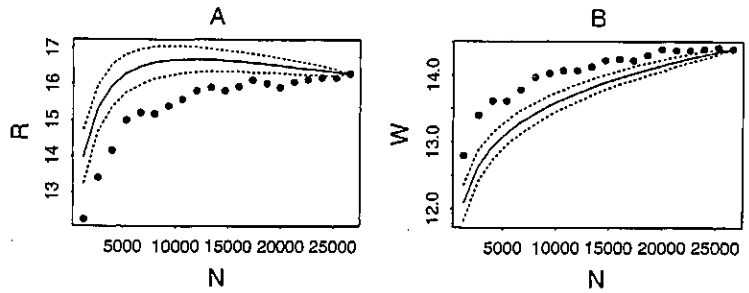

Figure 1.13: *The text characteristics R (panel A) and W (panel B) as a function of the sample size N in* Alice in Wonderland. *The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs.*

panel shows that W increases as a power function, $y = N^{0.514}$, a straight line in the double logarithmic plane. This example shows that once all types have been sampled, the constants systematically change when the sample size is increased. Because actual texts do not use all types that are available in the language, the changing magnitude of V affects the values of R and W. This is most clearly visible in the case of R. The left panel of Figure 1.13 shows that in *Alice's Adventures in Wonderland* R reveals the same downward curvature visible in Figure 1.14, but only after the first quarter of the text has been seen. Even for larger sample sizes, the appearance of new types slows down the rate at which R decreases, and gaurantees that its value stays well above 1.0 throughout the text.

The next two measures focus on the low-frequency words in the frequency spectrum. Sichel (1975) observed that the proportion S of dis legomena $V(2, N)$ of the vocabulary size $V(N)$ is more or less constant:

$$S = V(2, N)/V(N). \tag{1.19}$$

The proportion of hapax legomena on the vocabulary $V(1, N)/V(N)$ plays a key role in a measure proposed by Honoré (1979):

$$H = 100\frac{\log N}{1 - \frac{V(1,N)}{V(N)}}. \tag{1.20}$$

The idea underlying (1.20) is that $V(1, N)/V(N)$ is a linear function of $\log N$:

$$\frac{V(1, N)}{V(N)} = a - b \log N$$

(see panels A and B of Figure 1.15). Since for $N = 1$ the number of hapaxes is equal to the number of types, a must be equal to 1 and hence $b = 100/H$. It

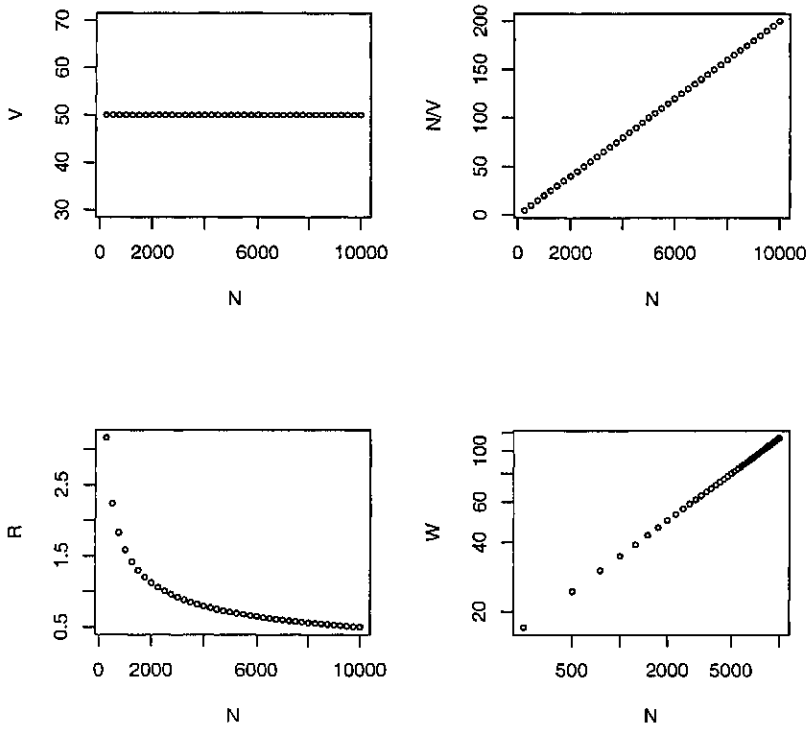

Figure 1.14: *V, N/V, R, and W as a function of N in a random sample of 10000 tokens from a population with 50 equiprobable types.*

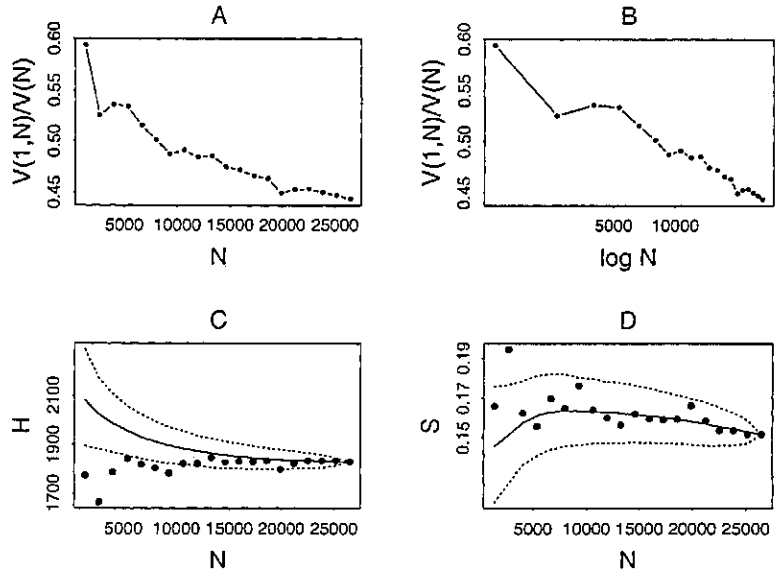

Figure 1.15: *The text characteristics H (panel C) and S (panel D) as a function of the sample size N in* Alice in Wonderland. *The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on 5000 permutation runs. Panels A and B show the roughly linear dependency of the relative number of hapax legomena on* $\log N$ *that underlies H.*

follows immediately that $H = 100/b$. Honoré does not consider the proportion $V(1, N)/V(N)$ by itself because this proportion generally decreases with increasing sample size (see panel B of Figure 1.15). The large dots in panel C of Figure 1.15 suggests that H is more or less constant for $N > 5000$ and converges rapidly to its final value of 1850. With respect to Sichel's constant S, the large dots in panel D reveal a slightly decreasing pattern with relatively large local fluctuations.

Herdan (1964) proposed a constant that is based on the observation that the growth curve of the vocabulary appears as roughly a straight line in the double logarithmic plane, as shown in panel A of Figure 1.16 for *Alice in Wonderland* by means of large dots. This suggests that a power model for $V(N)$ would be appropriate:

$$V(N) = aN^C.$$

Applying logarithms to both sides, we again obtain the equation for a straight line

$$\log V(N) = \log a + C \log N$$

with $\log a$ as intercept and C as slope. For sample size $N = 1$, $V(N)$ must also equal unity, and since $\log(1) = 0$, we have that $a = 1$. The remaining

looking-glass, it is in this very same text that K displays its greatest variability. In addition, the two texts reveal reasonably similar values for K for the first ten thousand tokens. Although *Through the looking-glass* has the lower overall repeat rate — note that this ties in nicely with its slightly higher lexical richness — the variability in the repeat rate itself within *Through the looking-glass* is so large that it becomes difficult to argue on the basis of K alone that *Through the looking-glass* differs from *Alice in Wonderland*. Texts are complex entities, and by using simple summary statistics one runs the risk of opting for too coarse a measure to identify similarities and differences between texts.

1.5 The lognormal distribution

An important model for word frequency distributions is the lognormal model. A random variable X is lognormally distributed if $Y = \log(X)$ follows a normal distribution. The lognormal model is sometimes used for skewed distributions with slowly decreasing right tails. Word frequency distributions are heavily skewed in this sense. Herdan (1960) and Carroll (1967) have therefore considered the possibility that word frequencies are lognormally distributed. The lognormal hypothesis is of special interest because many statistical tests presuppose normality. Word frequency distributions are decidedly non-normal. However, if they can be transformed into normal distributions by considering log frequency instead of absolute frequency, then these statistical tests can nevertheless be used after applying a simple logarithmic transformation.

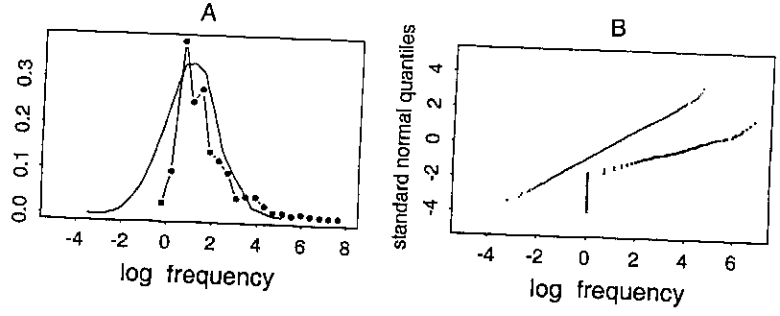

Figure 1.18: *The lognormal hypothesis. Panel A shows the estimated probability density function for log frequency in* Alice in Wonderland *(dashed and dotted line) and the estimated density function of a lognormal random variable with the same mean and standard deviation. Panel B plots the corresponding quantiles of the standard normal distribution.*

Unfortunately, it is not generally true that logarithmically transformed word frequencies are normally distributed. Figure 1.18 illustrates this point

for *Alice in Wonderland*. Panel A compares the distribution of log frequency in *Alice in Wonderland* (mean log frequency 0.974, variance 1.212, for 2651 word types) with 2651 random numbers from a lognormal distribution with the same logarithmic mean and variance by means of the estimated probability density functions. For the simulated lognormal distribution, represented by a solid line, we find a curve resembling the familiar bell-shaped curve of the normal distribution. The distribution of log frequency in *Alice in Wonderland*, represented by the dashed and dotted line, remains a skewed distribution.

Panel B is the corresponding Normal quantile-quantile plot. The horizontal axis plots log frequency sorted according to increasing frequencies. The vertical axis plots the quantiles of the standard normal, i.e., the values of a standard normal distribution corresponding to a given percentage of the sorted data. (For instance, the quantile for 2.5% of the data has the value of -1.96.) Normally distributed random variables show up as straight lines in Normal quantile-quantile plots, deviations from normality emerge as nonlinearities. In panel B, the upper line represents the simulated lognormal distribution, which, as expected, is a straight line. The log frequencies of *Alice in Wonderland* reveal a different pattern. Word frequencies are integer-valued, hence no log frequencies less than zero are attested. The vertical line at log frequency equals zero represents the hapax legomena, which jointly account for some 4.5% of all tokens. For the dis legomena and higher frequency ranks, the cumulated number of tokens reveals a more linear development.

Except for the lowest frequencies in the spectrum, the lognormal hypothesis seems to be reasonably well supported. This suggests that the problem that we are dealing with is at least in part one of discretization: Word frequency distributions are discrete. Normality, by contrast, presupposes a continuous random variable. Low-probability words either occur, in which case they are very likely to appear among the hapax legomena, or they do not occur. In the latter case, they do not appear in counts with fractional frequencies. Instead, they belong to the $V(0, N)$ unknown words with a frequency of zero.

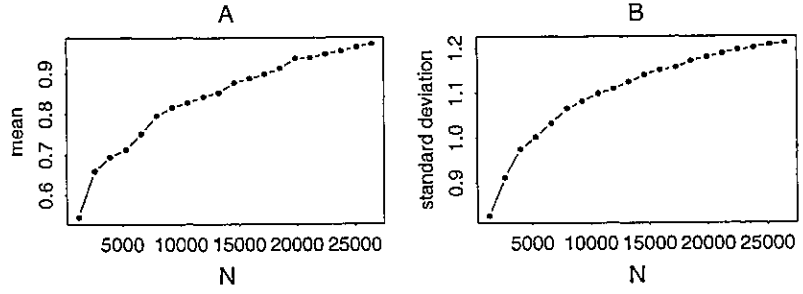

Figure 1.19: *The parameters of the lognormal model as a function of the sample size N for* Alice in Wonderland.

Even if we accept the lognormal model as a continuous approximation for

a discrete distribution, we again run into the problem that the parameters of the lognormal model change when the sample size is changed. Figure 1.19 illustrates this by now familiar phenomenon for *Alice in Wonderland*. Panel A plots the mean log frequency for twenty equally-spaced intervals, and panel B the corresponding standard deviations. Both the mean and the standard deviation appear as increasing functions of N. In chapter 3, we will show how the hypothesis of lognormality can be adjusted to avoid these problems.

1.6 Discussion

The main thrust of this chapter has been to show that for word frequency distributions the sample mean frequency and many other summary measures change in a highly systematic way as a function of the sample size. The parameters of the zeta (Zipf) and lognormal distributions are subject to exactly the same kind of systematic dependency on the sample size.

An additional complicating factor is that words are not randomly distributed in texts. Randomization tests show that the articles *a* and *the* are not uniformly distributed in *Alice in Wonderland*. Non-randomness in word usage similarly affects the measures that have been proposed as invariant with respect to sample size. Consequently, measures such as K and D, which in theory are true constants, nevertheless may reveal significantly non-random developmental profiles.

The prominent role of the sample size in shaping word frequency distributions, combined with the non-random way in which authors use their words in discourse, raises two important issues. First, when comparing texts or corpora, the characteristic constants that have been proposed as independent of the sample size should be interpreted with caution. They may reveal differences between authors, genre, or register, but when substantial differences in sample size are involved, the extent to which their values vary with sample size should be carefully considered. In addition, to gauge the importance of a difference in the value of a text characteristic for two or more texts, one should weigh the intertextual differences with respect to the intratextual variability of the text characteristic. It is only when the intertextual differences are larger than the intratextual differences that one may have some confidence that the differences are reliable.

Second, the law of large numbers does not appear to hold for word frequency distributions. Is the non-randomness in word use illustrated for the articles *the* and *a* in *Alice in Wonderland* to be held responsible? Or are lexical samples, even when encompassing tens of thousands or even millions of words, too small to allow the theoretical probabilities of words to be estimated from their sample relative frequencies? These issues are addressed in detail in the next chapters.

1.7 Bibliographical Comments

An elementary introduction to lexical statistics is Muller (1977). Muller (1979b) provides a useful collection of papers in the French tradition of lexical statistics. Other important studies in this tradition are Guiraud (1954), Brunet (1978), Honoré (1979), and Menard (1983). A more recent textbook is Lebart and Salem (1994).

In the Anglo-Saxon tradition, classic studies are Zipf (1935), Zipf (1949), Yule (1944), Carroll (1967), and Herdan (1960), Herdan (1964), Herdan (1966). Important technical papers are Good (1953), Good and Toulmin (1956), and Efron and Thisted (1976).

In the Eastern-European tradition, Orlov (1983a) and Orlov (1983b) are accessible studies. Guiter and Arapov (1983) is a useful collection of studies on Zipf's law. The concept of structural distributions is developed in Khmaladze and Chitashvili (1989) (in Russian), part of which has appeared in English in Khmaladze (1987). A review article in English is Chitashvili and Baayen (1993).

For Monte Carlo methods, see Hammersley and Handscomb (1964) and Meyer (1956). A review of lexical constants is Tweedie and Baayen (1998), for non-randomness in the use of function words, see Damerau (1975).

Important journals are *Computers and the Humanities, Literary and Linguistic Computing, Journal of Quantitative Linguistics, Computational Linguistics,* and the book series *Glottometrika*. A number of important technical papers can be found in *Biometrika*.

1.8 Questions

1. Show, using definition 1.2, that the ratio $N/V(N)$ represents the mean token frequency.

2. How might non-randomness in word usage affect the accuracy of theoretical estimates for $V(N)$?

3. Figure 1.20 plots the relative sample frequencies of *a* and *the* in *Through the looking-glass*. Offer an explanation for the developmental profiles.

4. Interpret the expression $V(0, N)$. What does Figure 1.3 suggest about its magnitude, when we view *Alice in Wonderland* as a sample of Carroll's word use?

5. Figure 1.21 plots the frequency spectrum of *Through the looking-glass* for $N = 14514$ and $N = 29028$. Does the curve with the greater number of hapax legomena ($V(1, N)$) represent the larger sample or the smaller one?

6. Rewrite
$$g(m, N) > 2g(m + 1, N) - g(m + 2, N)$$
in terms of the frequency spectrum $V(m, N)$.

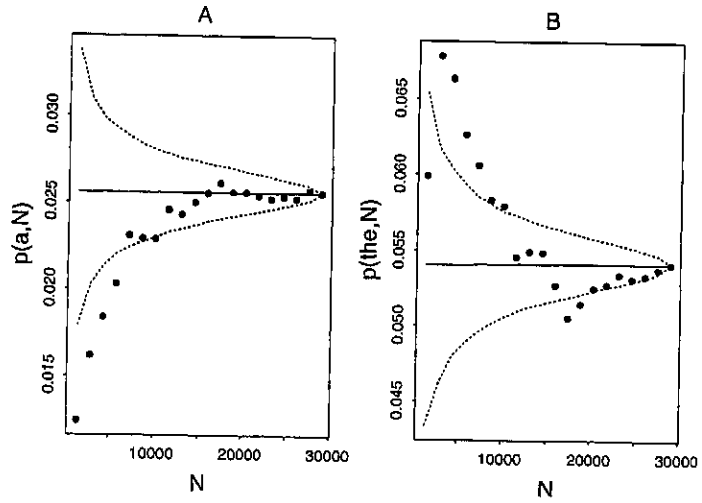

Figure 1.20: *The sample relative frequency of the article* a $p(a, N)$ *(panel A) and the sample relative frequency of the article* the $p(the, N)$ *(panel B) as a function of sample size* N *in* Through the looking-glass, *measured at 20 equally-spaced intervals. The large dots represent the empirical values, the solid line the Monte Carlo mean, and the dotted lines the 95% Monte Carlo confidence interval, based on a total of 5000 permutation runs.*

7. Figure 1.22 shows the error function of Zipf's zeta function for *Alice in Wonderland* at $N = 13250$, i.e., for each Zipf rank z it plots the difference between the observed frequency $f_z(z, 13250)$ and the expected frequency given by (1.7). What is the source of the striations at the right-hand side of the plot? Comment on the error pattern and its relevance for interpreting the high correlation ($r = -0.986, t_{(1827)} = -250.02, p = 0$) between Zipf rank and frequency.

8. Mandelbrot (1953) enriched Zipf's zeta distribution with a second free parameter b to enhance the model's accuracy for the high-frequency ranks:

$$f_z(z, N) = \frac{C}{(z + b)^a} .$$

   Express $V(m, N)$ as a function of m and the parameters of the Zipf-Mandelbrot distribution.

9. Figure 1.7 shows that the intercept is an increasing function of N and that the slope is a decreasing function of N. Offer an explanation for this pattern.

10. Rewrite K (1.15) in terms of the word frequencies $f(i, N)$ instead of in terms of the frequency spectrum $V(m, N)$.

11. The randomized version of *Alice in Wonderland* reveals higher values for C than the original non-randomized version. Is the direction of the difference between the two curves, higher values for the randomized version, what you would expect, or might the randomized version just as well have revealed lower values for C?

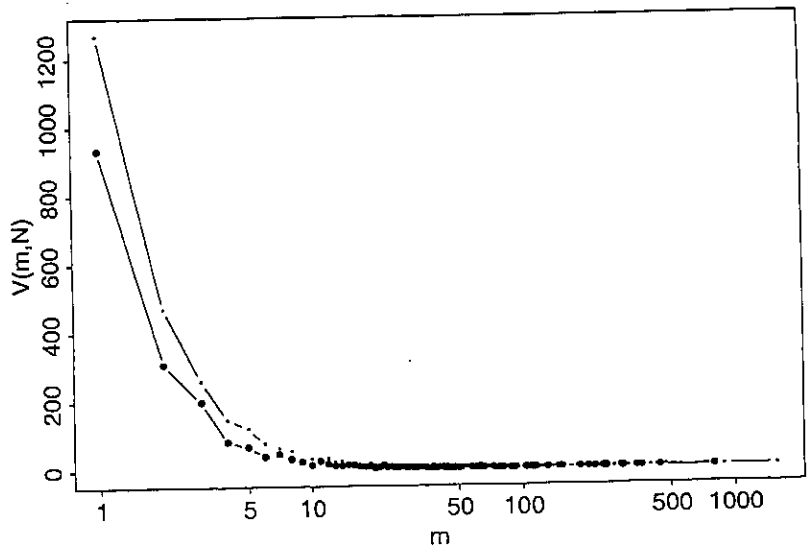

Figure 1.21: *The frequency spectra of* Through the looking-glass *at the sample sizes* $N = 14514$ *and* $N = 29028$ *(m: frequency class;* $V(m, N)$ *: number of types occurring* m *times).*

12. Figure 1.19 shows that the rate at which mean and standard deviation change as N is increased decreases. Under what circumstances do you expect the mean and the standard deviation to have a limit for $N \to \infty$?

13. Discuss the usefulness of the type-token ratio V/N for the evaluation of the lexical richness of texts of different lengths.

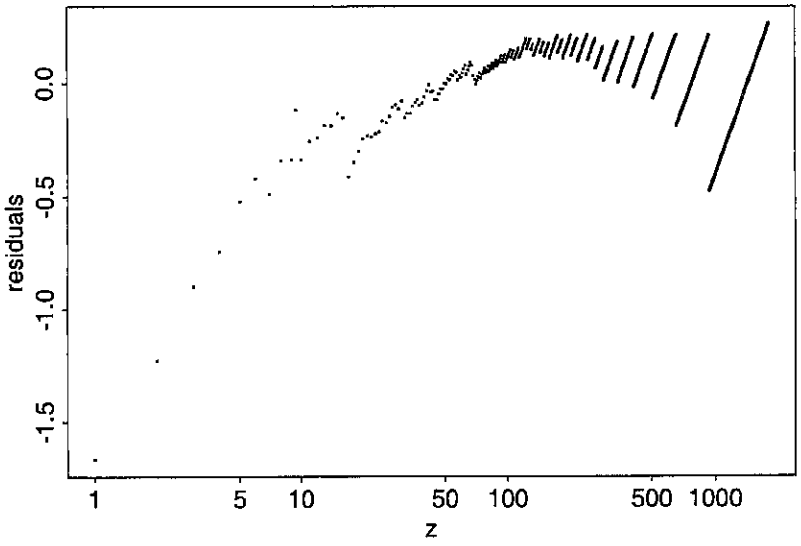

Figure 1.22: *The error function of Zipf's zeta function for* Alice in Wonderland *at* $N = 13250$.

Chapter 2

Non-parametric models

This chapter presents a range of statistical techniques that are available for the analysis of word frequency distributions. Section 2.1 introduces some basic probabilistic concepts. Section 2.2 discusses the urn model, according to which word use is viewed as random selection from a population with fixed probabilities for words to occur. The binomial model and the Poisson approximation to the binomial model are defined here. Section 2.3 is concerned with the structural type distribution, which allows us to restate the Poisson model in integral form. Section 2.4 introduces the concept of the LNRE zone, the range of sample sizes where the sample relative frequencies are not good estimates of the corresponding population probabilities. The next section (2.5) focuses on the Good-Turing estimates, which adjust sample relative frequencies for the non-negligible frequency weight of the unseen words. Methods for calculating the frequency spectrum for any sample size given the frequency spectrum for a given sample size are presented in sections 2.6 and 3.2.

2.1 Basic concepts

SUMMARY *This section briefly presents the definitions of the expectation, variance, and covariance of a random variable, in combination with a summary of some basic properties of these operators. The Bernouilli distribution is also introduced, and the distinction between permutations and combinations is outlined.*

In this section a summary review is presented of the main definitions and basic properties that we will need in the course of this chapter. For detailed discussion and proofs, the reader is referred to textbooks such as, e.g., Ross (1988).

Definition 2.1 The **expectation** $\mathrm{E}[X]$ of a random variable X is given by

$$\mathrm{E}[X] = \sum_x x \Pr(X = x).$$