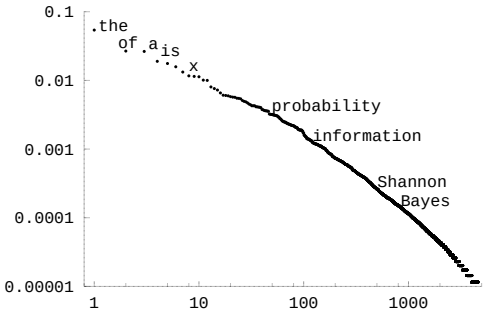

Figure 18.5. Log–log plot of frequency versus rank for the words in the LaTeX file of this book.

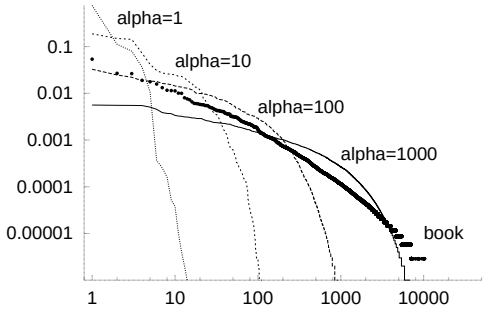

Figure 18.6. Zipf plots for four 'languages' randomly generated from Dirichlet processes with parameter α ranging from 1 to 1000. Also shown is the Zipf plot for this book.

The Dirichlet process

Assuming we are interested in monogram models for languages, what model should we use? One difficulty in modelling a language is the unboundedness of vocabulary. The greater the sample of language, the greater the number of words encountered. A generative model for a language should emulate this property. If asked 'what is the next word in a newly-discovered work of Shakespeare?' our probability distribution over words must surely include some non-zero probability for *words that Shakespeare never used before*. Our generative monogram model for language should also satisfy a consistency rule called *exchangeability*. If we imagine generating a new language from our generative model, producing an ever-growing corpus of text, all statistical properties of the text should be homogeneous: the probability of finding a particular word at a given location in the stream of text should be the same everywhere in the stream.

The Dirichlet process model is a model for a stream of symbols (which we think of as 'words') that satisfies the exchangeability rule and that allows the vocabulary of symbols to grow without limit. The model has one parameter α. As the stream of symbols is produced, we identify each new symbol by a unique integer w. When we have seen a stream of length F symbols, we define the probability of the next symbol in terms of the counts $\{F_w\}$ of the symbols seen so far thus: the probability that the next symbol is a new symbol, never seen before, is

$$\frac{\alpha}{F+\alpha}. \tag{18.11}$$

The probability that the next symbol is symbol w is

$$\frac{F_w}{F+\alpha}. \tag{18.12}$$

Figure 18.6 shows Zipf plots (i.e., plots of symbol frequency versus rank) for million-symbol 'documents' generated by Dirichlet process priors with values of α ranging from 1 to 1000.

It is evident that a Dirichlet process is not an adequate model for observed distributions that roughly obey Zipf's law.

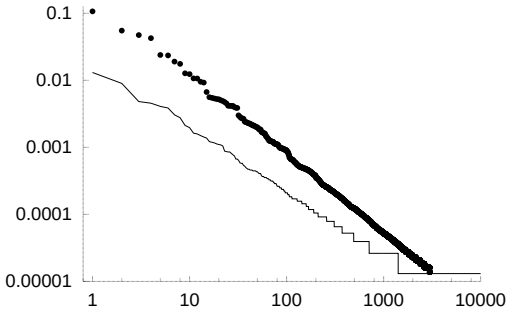

Figure 18.7. Zipf plots for the words of two 'languages' generated by creating successive characters from a Dirichlet process with $\alpha = 2$, and declaring one character to be the space character. The two curves result from two different choices of the space character.

With a small tweak, however, Dirichlet processes can produce rather nice Zipf plots. Imagine generating a language composed of elementary symbols using a Dirichlet process with a rather small value of the parameter α, so that the number of reasonably frequent symbols is about 27. If we then declare one of those symbols (now called 'characters' rather than words) to be a space character, then we can identify the strings between the space characters as 'words'. If we generate a language in this way then the frequencies of words often come out as very nice Zipf plots, as shown in figure 18.7. Which character is selected as the space character determines the slope of the Zipf plot – a less probable space character gives rise to a richer language with a shallower slope.

▶ 18.3 Units of information content

The information content of an outcome, x, whose probability is $P(x)$, is defined to be

$$h(x) = \log \frac{1}{P(x)}. \tag{18.13}$$

The entropy of an ensemble is an average information content,

$$H(X) = \sum_x P(x) \log \frac{1}{P(x)}. \tag{18.14}$$

When we compare hypotheses with each other in the light of data, it is often convenient to compare the log of the probability of the data under the alternative hypotheses,

$$\text{'log evidence for } \mathcal{H}_i\text{'} = \log P(D \,|\, \mathcal{H}_i), \tag{18.15}$$

or, in the case where just two hypotheses are being compared, we evaluate the 'log odds',

$$\log \frac{P(D \,|\, \mathcal{H}_1)}{P(D \,|\, \mathcal{H}_2)}, \tag{18.16}$$

which has also been called the 'weight of evidence in favour of $\mathcal{H}_1$'. The log evidence for a hypothesis, $\log P(D \,|\, \mathcal{H}_i)$ is the negative of the information content of the data D: if the data have large information content, given a hypothesis, then they are surprising to that hypothesis; if some other hypothesis is not so surprised by the data, then that hypothesis becomes more probable. 'Information content', 'surprise value', and log likelihood or log evidence are the same thing.

All these quantities are logarithms of probabilities, or weighted sums of logarithms of probabilities, so they can all be measured in the same units. The units depend on the choice of the base of the logarithm.

The names that have been given to these units are shown in table 18.8.