Chapter 38: Statistical methods for corpus exploitation

Marco Baroni (baroni@sslmit.unibo.it)
SITLEC, University of Bologna
Corso Diaz 64, 47100 Forlì, Italy

Stefan Evert (stefan.evert@uos.de)
Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany

September 15, 2006

Contents

1 Introduction

Linguists look for generalizations and explanations of various kinds of linguistic phenomena. While the interest is usually in an *intensional* view of these phenomena, to be explained in terms of the human language competence, such competence cannot be directly observed. Thus, evidence has to come from an external reflection of it, i.e., it has to be based on an *extensional* view of language. According to this extensional view, a language is defined as the set of all utterances produced by speakers of the language (with all the paradoxes that this view implies – see, e.g., Chomsky 1986, chapter 2). Corpora are finite samples from the infinite set that constitutes a language in this extensional sense. For example, in this perspective, the Brown corpus (see Article 22) is a finite sample of all the utterances produced in written form by American English speakers. Psycholinguistic experiments, such as eye-tracking tests, priming, and even traditional grammaticality judgments (Schütze 1996) constitute other sources of evidence. It is important to observe that the empirical analysis of these other sources also requires an extensional view of language.

It is rarely the case that linguists are interested in the samples *per se*, rather than in generalizations from the samples to the infinite amount of text corresponding to the extensional definition of a (sub)language. For example, a linguist studying a pattern in the 500 text samples of the Brown corpus will typically be interested in drawing conclusions about (written) American English as a whole, and not just about the specific texts that compose the Brown. Statistical inference allows the linguist to generalize from properties observed in a specific sample (corpus) to the same properties in the language as a whole (statistical inference, on the other hand, will not be of help in solving thorny issues such as what is the appropriate extensional definition of a "language as a whole" and how we can sample from that).

Statistical inference requires that the problem at hand is *operationalized* in quantitative terms, typically in the form of units that can be *counted* in the available sample(s). This is the case we will concentrate on here (but see Section 8 for other kinds of measurements). For example, a linguist might be interested in the issue of whether a certain variety of English is more "formal" than another (as in some of Douglas Biber's work, see Article 40). In order to operationalize this research question, the linguist might decide to take passivization as a cue of formality, and count the number of sentences that display passivization in samples from the two varieties. Statistical inference can then be used to generalize from the difference in number of passives between the two samples to the difference between the two varieties that the samples represent (we will discuss this example and the appropriate techniques further in Section 5). Similarly, a linguist might be interested in whether (certain classes of) idiomatic constructions have a tendency to repel passive formation (as observed by Culicover/Jackendoff 2005 and many others). In order to operationalize this question, the linguist may count the number of passives in idiomatic and non-idiomatic phrases in a corpus. Statistical inference will then help to determine how reliably the attested difference in passive frequency would generalize to idiomatic and non-idiomatic phrases in general. Of course, it is up to the linguist to interpret the generalizations about frequencies produced by statistical analysis in terms of the linguistic phenomena of interest.

Statistical inference is necessary because any sample from a language is subject to some random variation. Suppose that someone doubted the claim that non-idiomatic constructions are more prone to passivization than idiomatic constructions, and we wanted to dispel these doubts. A sample of language that reveals a higher proportion of passives among the non-idiomatic constructions, especially if the difference in proportions is small, would not allow us to reject the doubters' hypothesis: even if they were right, we could not expect the proportions to be exactly identical in *all* samples of language. Statistical inference can help us to determine to what extent the difference between a sample-based observation and a theoretical prediction can be taken as serious evidence that the prediction made by the theory is wrong, and to what extent it can reasonably be attributed to random variation. In the case at hand, statistical inference would tell us whether the difference in passive rates in the two samples can be explained by random variation, or whether it is the symptom of a true underlying difference. It is perhaps worth clarifying from the outset that randomness due to sampling has to be distinguished from measurement errors, such as those introduced by the automatic annotation and analysis of corpus data (something that statistical methods will not help us correct). Suppose that a very skilled linguist sampled 100 English sentences and recorded very carefully how many of them are passives, without making any error. It should be intuitive that, given another random sample of 100 sentences and the same error-free linguist, the exact number of passives would probably be different from the one found in the previous sample. This is the random variation we are referring to here.

Notice that the necessity of statistical inference pertains to the need to generalize from a finite (random) sample of language data to the theoretically infinite amount of text corresponding to the extensional definition of an entire (sub)language, and it has nothing to do with whether our theory about the phenomenon at hand, or about language competence in general, includes a probabilistic component. The prediction that idiomatic sentences repel the passive construction might stem from a completely categorical theory of how passives and idiomaticity interact – still, randomly sampled English sentences will display a certain amount of variation in the exact proportion of passives they contain.

The rest of this article introduces the basics of statistical inference. We use the artificially simple example of testing a hypothesis about the proportion of passives in English sentences (and later proportions of passives in sentences from different English genres), in order to focus on the general philosophy and methodology of statistical inference as applied to corpus linguistics, rather than on the technical details of carrying out the relevant computations, which can be found in many general books on the subject and are implemented in all standard statistical packages (see references in Section 9). Section 6 gives examples of how statistical inference can be applied to more realistic linguistic analysis settings.

2 The logic behind hypothesis testing

Imagine that an American English style guide claims that 15% of the sentences in the English language are in the passive voice (as of June 2006, `http://www.ego4u.com/en/business-english/grammar/passive` makes the even bolder statement that no more than 10% of English sentences are in the passive voice and writers should be careful to use passives sparingly). This is a fairly easy claim to operationalize, since it is already phrased in terms of a proportion. However, we still need to define what we understand by "the English language", and what it means for a sentence to be in the passive voice. Given the source of the claim and our need for an extensional definition, it makes sense to take "English" to mean the set of all English texts published in the US and produced by professional writers. Regarding the second issue, we consider a sentence to be in the passive voice if it contains at least one verb in the passive form, which seems to be a plausible interpretation of what the style guide means (after all, it is warning against the overuse of passives), and at the same time makes it easier to count the number of sentences in passive voice using automated pattern matching techniques (which might not be relevant with the small samples we use here, but would be important when dealing with large amounts of data).

It is of course impossible to look at all sentences in all the publications satisfying the criteria above – what we can do, at best, is to select a random sample of them. In particular, we took a random sample of 100 sentences of the relevant kind, and we counted the number of them containing a passive. For convenience, we restricted ourselves to publications from 1961, because we are lucky enough to already own a random sample of sentences of the relevant kind from that year – namely, the Brown corpus! All we had to do was select 100 random sentences from this random sample (we will see in Section 7 that it is not entirely correct to treat sentences from the Brown as a random sample, but we ignore this for now).

If the style guide's claim is true, we would expect 15 sentences to be in the passive voice. Instead, we found 19 passives. This seems to indicate that the proportion is higher than 15% and rather close to 20%. However, it is obvious that, even if the claim of the style guide was correct, not *all* samples of size 100 would have exactly 15 passives, because of random variation. In light of this, how do we decide whether 19 passives are enough to reject the style guide's claim?

In statistical terms, the claim that we want to verify is called a *null hypothesis*, $H_0 : \pi = 15\%$, where π is the putative proportion of passives in the set of sentences that constitute our extensional definition of American English. This set of sentences is usually called a *population* in statistical parlance, and the goal of statistical inference is to draw conclusions about certain properties of this population from an available sample (the population itself is practically infinite for all intents and purposes, and we can only access a small finite subset of it). We will often refer to π as a population proportion or parameter in what follows. The number of sentences we have randomly sampled from the population is called the *sample size*, $n = 100$. Intuitively, we expect $e = n \cdot \pi = 15$ passives in the sample if the null hypothesis is true. This is called the *expected frequency.* The number of passives we actually observed in the sample, $f = 19$, is called the *observed frequency.*

Having introduced the terminology, we can rephrase the problem above as follows. If we are prepared to reject the null hypothesis that $\pi = 15\%$ for an observation of $f = 19$, there is a certain risk that in doing so we are making the wrong decision. The question is how we can quantify this risk and decide whether it is an acceptable risk to take. Imagine that the null hypothesis in fact holds, and that a large number of linguists perform the same experiment, sampling 100 sentences and counting the passives. We can then formally define risk by the percentage of linguists who wrongly reject the null hypothesis, and thus publish incorrect results. In particular, if our observation of $f = 19$ is deemed sufficient for rejection, all the other linguists who observed 19 or even more passives in their samples would also reject the hypothesis. The risk is thus given by the percentage of samples containing 19 or more passives that would be drawn from a language in which the true proportion of passives is indeed 15%, as stipulated by H_0. Rejecting the null hypothesis when it is in fact true is known as a *type-1 error* in the technical literature (failure to reject H_0 when it does not hold constitutes a *type-2 error*, which we do not discuss here, but see, e.g., DeGroot/Schervish 2002, Chapter 8).

Fortunately, we do not need to hire hundreds of linguists to compute the risk of wrong rejection, since the thought experiment above is fully equivalent to drawing balls from an urn. Each ball represents a sentence of the language, with red balls for passive sentences and white balls for sentences in other voices. The null hypothesis stipulates that the proportion of red balls in the urn is 15%. The observed number of red balls (passives) changes from sample to sample. In statistical terminology, it is called a *random variable*, typically denoted by a capital letter such as X. We simulate a large number of samples from the urn with a computer and tabulate how often each possible value k of the random variable X is observed. The result of this simulation is shown in Figure 1, which reports percentages of samples that yield $X = k$ for k ranging from 0 to 30 (the percentage is indistinguishable from 0 for all values outside this range). For instance, the value $X = 19$ can be observed in 5.6% of the samples. The information presented in this graph is called the *sampling distribution* of X under H_0. The percentage of samples with $X = k$ is called the *probability* $\Pr(X = k)$. For example, $\Pr(X = 19) = 5.6\%$ (our reasoning in this Section has led us to what is known as the *frequentist* definition of probabilities; we do not discuss the alternative *Bayesian* interpretation of probability theory here, but see for example Section 1.2 of DeGroot/Schervish 2002).

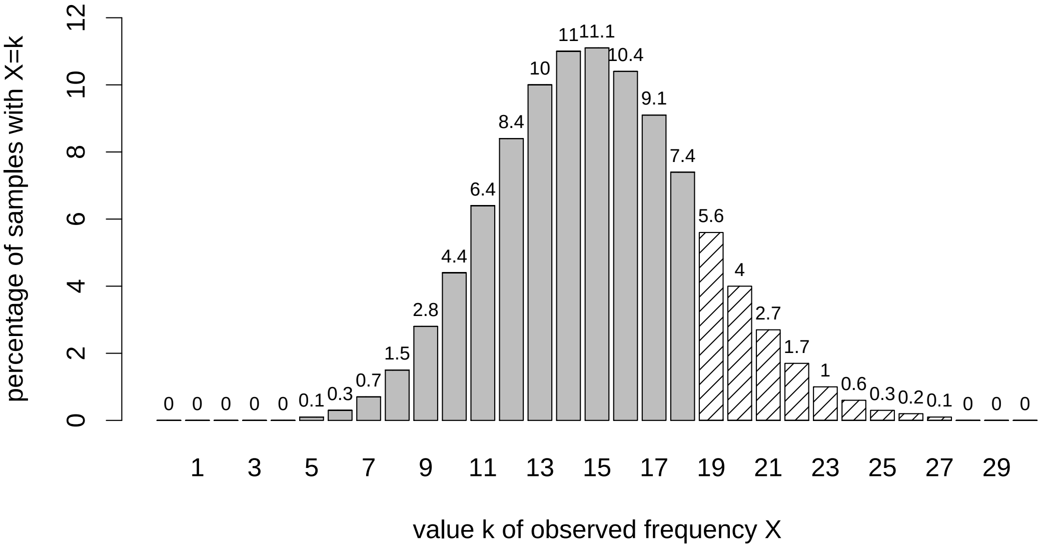

Figure 1: Sampling distribution of X with $n = 100$ and $\pi = 15\%$.

Following the discussion above, the risk of wrongly rejecting the null hypothesis for an observation of $f = 19$ is given by the percentage of samples with $X \geq 19$ in the sampling distribution, i.e., the probability $\Pr(X \geq 19)$. This probability can be computed by summing over the shaded bars in Figure 1:

$$\Pr(X \geq 19) = \Pr(X = 19) + \Pr(X = 20) + \cdots + \Pr(X = 100) = 16.3\% \qquad (1)$$

This is called a *tail probability* because it sums over the right-hand "tail" of the distribution. In the same way, we can compute the risk associated with any other value f, namely the probability:

$$\Pr(X \geq f) := \sum_{k=f}^{n} \Pr(X = k) \qquad (2)$$

We refer to this risk as the *p-value* of an observation f. Notice that *smaller* p-values are *more* significant, since they indicate that it is less risky to reject the null hypothesis, and hence they allow greater confidence in the conclusion that the null hypothesis should be rejected. In our example, the p-value of $f = 19$ indicates that the risk of false rejection is unacceptably high at $p = 16.3\%$. If we used $f = 19$ as the threshold for rejection and the null hypothesis happened to be true, about one in six experiments would lead to wrong conclusions. In order to decide whether to reject H_0 or not, the computed p-value is often compared with a conventional scale of *significance levels*. A p-value below 5% is always required to consider a result significant. Other common significance levels are 1% and 0.1% (usually written as mathematical fractions rather than percentages and denoted by the symbol α, viz. $\alpha = .05$, $\alpha = .01$ and $\alpha = .001$).

So far, we have only been considering cases in which the observed frequency is greater than the one predicted under H_0 – reflecting our intuition that the proportion of passives proposed by the style guide errs on the side of being too low, rather than too high. However, in principle it would also be possible that the proportion of passives is *lower* than predicted by H_0. Coming back to our passive-counting linguists, if they are prepared to reject H_0 for $f = 19$, they should also reject it for $X = 7$ or $X = 8$, since these values are even more "extreme" than 19 with respect to H_0. Thus, when computing the p-value of $f = 19$, it is typically appropriate to sum over the probabilities of all values that are at least as "extreme" as the observed value, to either side of the expected frequency e, since they add to the risk of false rejection.

It is difficult to determine exactly which of the values below e should count as equally extreme as, or more extreme than f, but one reasonable approach is to include all the values of X with an absolute difference $|X - e| \geq |f - e|$. Using $|f - e|$ (or, more precisely, $(f - e)^2$, which has certain mathematical advantages) as a measure of "extremeness" leads to a class of statistical tests known as *chi-squared tests.* An alternative approach would rather compare the probabilities $\Pr(X = k)$ and $\Pr(X = f)$ as a measure of extremeness, resulting in a class known as likelihood (ratio) tests. In many common cases, both classes of tests give very similar results. We will focus on chi-squared tests in this article, but see Article 57 for an application where likelihood tests are known to be superior.

In the case at hand, using the chi-squared criterion, the p-value would be computed by adding up the probabilities of $X \geq 19$ and $X \leq 11$ (since $|19 - 15| = 4 = |11 - 15|$). In the illustration shown in Figure 1, we would add the bars for $k = 1 \ldots 11$ to the shaded area. This way of computing the p-value, taking both "extreme" tails of the distribution into account, is called a *two-tailed test* (the approach above, where we considered only one side, is known as a *one-tailed test*). Of course, a two-tailed p-value is always greater than (or equal to) the corresponding one-tailed p-value. In our running example, the two-tailed p-value obtained by summing over the bars for $X \leq 11$ and $X \geq 19$ turns out to be 32.6%, indicating a very high risk if we chose to reject the null hypothesis for $f = 19$ (you might obtain a different p-value if the binomial test implemented in your software package uses the likelihood criterion, although the value will still indicate a very high risk in case of rejection). Had our experiment yielded 22 passives instead, the one-tailed test would have produced a p-value of 3.9%, while the two-tailed test would have given a p-value of 6.7%. Thus, by adopting the common 5% significance threshold, we would have had enough evidence to reject the null hypothesis according to the one-tailed test, but not enough according to the two-tailed test.

As a general rule, one should always use the more conservative two-tailed test, unless there are very strong reasons to believe that the null hypothesis could only be violated in one direction – but it is hard to think of linguistic problems where this is the case (in many situations we can predict the probable direction of the violation, but there are very few cases where we would be ready to claim that a violation in the other direction is absolutely impossible). If we use a two-tailed test, the interpretation of a significant result will of course have to take into account whether f is greater or smaller than e. Observed frequencies of 25 and 5 passives, respectively, would both lead to a clear rejection of the null hypothesis that 15% of all sentences are in the passive voice, but they would require rather different explanations.

Not only have we been spared the expense of hiring passive-counting linguists to repeat the experiment; it is not even necessary to perform expensive computer simulation experiments in order to carry out the sort of tests we just illustrated, because $\Pr(X = k)$ – the percentage of samples of size n from a population with proportion π of passive sentences that would result in a certain value k of the random variable X – can be computed with the following formula, known as the *binomial distribution* (the hypothesis test we have described above, unsurprisingly, is called a *binomial test*):

$$\Pr(X = k) = \binom{n}{k} (\pi)^k (1 - \pi)^{n-k} \tag{3}$$

The binomial coefficient $\binom{n}{k}$, "n choose k", represents the number of ways in which an unordered set of k elements can be selected from n elements. Any elementary textbook on probability theory or statistics will show how to compute it; see, e.g., DeGroot/Schervish (2002, Section 1.8). Of course, all statistical software packages implement binomial coefficients and the binomial distribution.

For a different null hypothesis about the population proportion π or a different sample size n, we obtain sampling distributions with different peaks and shapes – in statistical terminology, π and n are the *parameters* of the binomial distribution. In particular, the value of π affects the location of the peak in the histogram. For example, if we hypothesized that $\pi = 30\%$, we would expect a peak around the expected value $e = n \cdot \pi = 30$ in the histogram corresponding to Figure 1. Intuitively, experiments in which we draw 1,000 balls will tend to produce outcomes that are closer to the expected value than experiments in which we draw 100 balls. Thus, by decreasing or increasing n, we obtain distributions that have narrower or wider shapes, respectively. A sample of size 100 is small by the standards of statistical inference. As Karl Pearson, one of the founding fathers of modern statistics, once put it: "Only naughty brewers deal in small samples!" (cf. Pearson 1990, p. 73; this quip was a reference to W. S. Gosset, an employee of the Guinness brewery who developed and published the now famous t-test under the pseudonym of "Student"). It will typically be difficult to reject H_0 based on such a sample, unless the true proportion is very far away from the null hypothesis, exactly because a small sample size leads to a wide sampling distribution. Had we taken a sample of 1,000 sentences and counted 190 passives, the null hypothesis would have been clearly rejected (a two-sided binomial test with $f = 190$, $n = 1000$ and $H_0 : \pi = 15\%$ gives a p-value of $p = 0.048\%$, sufficient for rejection even at the very conservative significance level $\alpha = .001$).

The procedure of hypothesis testing that we introduced in this section is fundamental to understanding statistical inference. At the same time, it is not entirely intuitive. Thus, before we move on, we want to summarize its basic steps. For the whole process to be meaningful, we must have a *null hypothesis* H_0 that operationalizes a research question in terms of a quantity that can be computed from observable data. In our case, the null hypothesis stipulates that the proportion of passives in the *population* of (professionally written American) English sentences is 15%, i.e.: $H_0 : \pi = 15\%$. We draw a random *sample* of size n of the relevant units (100 sentences in our case) from the population, and count the number of units that have the property of interest (in our case, being passive sentences). Given the population proportion stipulated by the null hypothesis and the sample size, we can determine a *sampling distribution* (by simulation or using a mathematical formula). The sampling distribution specifies, for each possible outcome of the experiment (expressed by the *random variable* X, which in our case keeps track of the frequency of passives in the sample), how likely it is under the null hypothesis. This *probability* is given by the percentage of a large number of experiments that would produce the outcome X in a world in which the null hypothesis is in fact true. The sampling distribution allows us, for every possible value k of X, to compute the *risk* of making a mistake when we are prepared to reject the null hypothesis for $X = k$. This risk, known as the *p-value* corresponding to k, is given by the overall percentage of experiments that give an outcome at least as extreme as $X = k$ in a world in which the null hypothesis is true (see above for the *one-* and *two-tailed* ways to interpret what counts as "extreme"). At this point, we look at the actual outcome of the experiment in our sample, i.e., the *observed* quantity f (in our case, f is the number of passives in a sample of 100 sentences), and we compute the p-value (risk) associated with f. In our example, the (two-tailed) p-value is 32.6%, indicating a rather high risk in rejecting the null hypothesis. We can compare the p-value we obtained with conventional thresholds, or *significance levels*, that correspond to "socially acceptable" levels of risk, such as the 5% threshold $\alpha = .05$. If the p-value is higher than the threshold, we say that the results of the experiment are not *statistically significant*, i.e., there is a non-negligible possibility that the results would be obtained by chance even if the null hypothesis is true.

Notice that a non-significant result simply means that our evidence is not strong enough to reject the null hypothesis. It does *not* tell us that the null hypothesis is correct. In our example, although the observed frequency is not entirely unlikely under the null hypothesis of a passive proportion of 15%, there are many other hypotheses under which the same result would be even more likely, most obviously, the hypothesis that the population proportion is 19%. Because of this indirect nature of statistical hypothesis testing, problems undergoing statistical treatment are typically operationalized in a way in which the null hypothesis is "uninteresting", or contradicts the theory we want to support. Our hope is that the evidence we gather is strong enough to reject H_0. We will come back to this in Section 5 below, presenting a two-sample setting where this strategy should sound more natural.

While many problems require more sophisticated statistical tools than the ones described in this section, the basic principles of hypothesis testing will be exactly the same as in the example we just discussed.

3 Estimation and effect size

Suppose that we ran the experiment with a sample of n = 1,000 sentences, $f = 190$ of which turned out to be in the passive voice. As we saw in the previous Section, this result with the larger sample leads to a clear rejection of the null hypothesis $H_0 : \pi = 15\%$. At this point, we would naturally like to know what the *true* proportion of passives is in edited American English. Intuitively, our best guess is the observed proportion of passives in the sample, i.e., $\hat{\pi} = f/n$. This intuitive choice can also be justified mathematically. It is then known as a *maximum-likelihood estimate* or *MLE* (DeGroot/Schervish 2002, Section 6.5).

Since we have estimated a single value for the population proportion, $\hat{\pi}$ is called a *point estimate*. The problem with point estimates is that they are subject to the same amount of random variation as the observed frequency on which they are based: most linguists performing the same experiment would obtain a different estimate $\hat{\pi} = X/n$ (note that, mathematically speaking, $\hat{\pi}$ is a random variable just like X, which assumes a different value for each sample).

Let us put the question in a slightly different way: besides the point estimate $\hat{\pi} = 19\%$, which other values of π are also plausible given our observation of $f = 190$ passives in a sample of n = 1,000 sentences? Since $H_0 : \pi = 15\%$ was rejected by the binomial test, we know for instance that the value $\pi = 15\%$ is *not* plausible according to our observation. This approach allows us to answer the question in an indirect way. For any potential estimate $\pi = x$, we can perform a binomial test with the null hypothesis $H_0 : \pi = x$ in order to determine whether the value x is plausible (H_0 cannot be rejected at the chosen significance level α) or not (H_0 can be rejected). Note that failure to reject H_0 does not imply that the estimate x is very likely to be accurate, but only that we cannot rule out the possibility $\pi = x$ with sufficient confidence. Figure 2 illustrates this procedure for six different values of x, when $f = 190$ and $n = 1,000$. As the figure shows, $H_0 : \pi = 17\%$ would not be rejected, and thus 17% is in our set of plausible values. On the other hand, $H_0 : \pi = 16.5\%$ would be rejected, and thus 16.5% is not in our set.

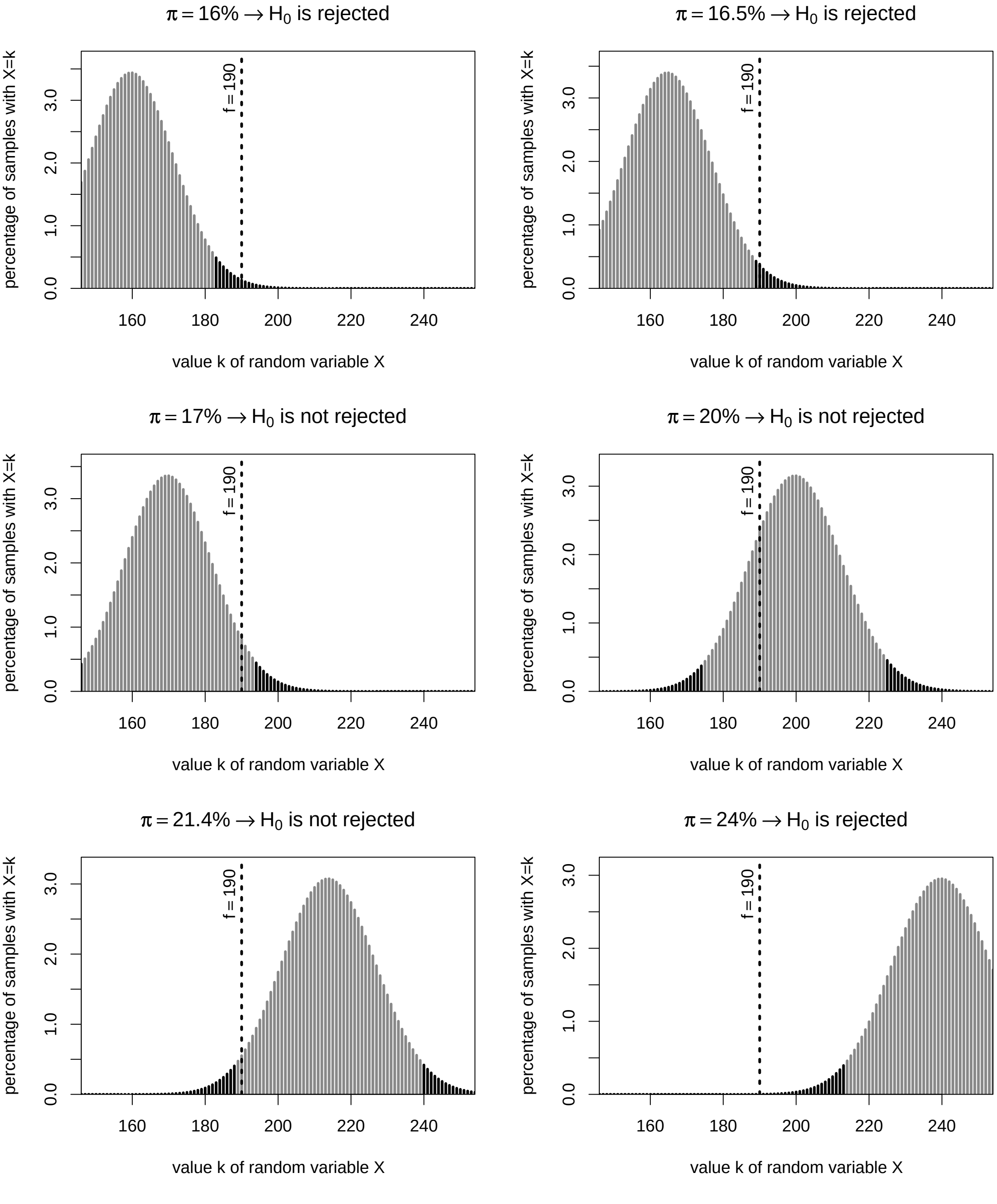

Figure 2: Illustration of the procedure for estimating a confidence set

	$n = 100$ $k = 19$	$n = 1{,}000$ $k = 190$	$n = 10{,}000$ $k = 1{,}900$
$\alpha = .05$	11.8% . . . 28.1%	16.6% . . . 21.6%	18.2% . . . 19.8%
$\alpha = .01$	10.1% . . . 31.0%	15.9% . . . 22.4%	18.0% . . . 20.0%
$\alpha = .001$	8.3% . . . 34.5%	15.1% . . . 23.4%	17.7% . . . 20.3%

Table 1: Binomial confidence intervals for various sample sizes n and confidence levels α. The maximum-likelihood estimate is $\hat{\pi} = 19\%$ in each case.

Collecting all plausible values $\pi = x$, we obtain a *confidence set*. For the binomial test, this confidence set is an uninterrupted range of numbers and is called a *binomial confidence interval*. Of course, it is infeasible to perform separate hypothesis tests for the infinite number of possible null hypotheses $\pi = x$, but specialized mathematical algorithms (available in all standard statistical software packages) can be used to compute the end points of binomial confidence intervals efficiently. In our example, the observed data $f = 190$ and $n = 1{,}000$ yield a confidence interval of $\pi \approx 16.6\% \ldots 21.6\%$ (the common mathematical notation for such a range, which you may encounter in technical literature, is $[.166, .216]$).

The width of a binomial confidence interval depends on the sample size n and the significance level α used in the test. As we have seen in Section 2, a larger value of n makes it easier to reject the null hypothesis. Obviously, adopting a higher (i.e., less conservative) value of α also makes it easier to reject H_0. Hence these factors lead to a narrower confidence interval (which, to reiterate this important point, consists of all estimates x for which H_0 is *not* rejected). Table 1 shows confidence intervals for several different sample sizes and significance levels. A confidence interval for a significance level of $\alpha = .05$ (which keeps the risk of false rejection below 5%) is often called a 95% confidence interval, indicating that we are 95% certain that the true population value π is somewhere within the range (since we can rule out any other value with 95% certainty). Similarly, a significance level of $\alpha = .01$ leads to a 99% confidence interval.

Confidence intervals can be seen as an extension of hypothesis tests. The 95% confidence interval for the observed data immediately tells us whether a given null hypothesis $H_0 : \pi = x$ would be rejected by the binomial test at significance level $\alpha = .05$. Namely, H_0 is rejected if and only if the hypothesized value x does *not* fall within the confidence interval. The width of a confidence interval illustrates thus how easily a null hypothesis can be rejected, i.e., it gives an indication of how much the (unknown) true population proportion π must differ from the value stipulated by the null hypothesis (which is often denoted by the symbol π_0) so that H_0 will reliably be rejected by the hypothesis test. Intuitively speaking, the difference between π and π_0 has to be considerably larger than the width of one side of the 95% confidence interval so that it can reliably be detected by a binomial test with $\alpha = .05$ (keep in mind that, even when the difference between π and π_0 is larger than this width, because of sampling variation, $\hat{\pi}$ and π_0 might be considerably closer, leading to failure to reject π_0). The term *effect size* is sometimes used as a generic way to refer to the difference between null hypothesis and true proportion. The reliability of rejection given a certain effect and sample size is called the *power* of the hypothesis test (see DeGroot/Schervish 2002, Chapter 8). In our example, the arithmetic difference $\pi - \pi_0$ is a sensible way of quantifying effect size, but many other measures exist and may be more suitable in certain situations (we will return to this issue during the discussion of two-sample tests in Section 5).

In corpus analysis, we often deal with very large samples, for which confidence intervals will be extremely narrow, so that a very small effect size may lead to highly significant rejection of H_0. Consider the following example: Baayen (2001, p. 163) claims that the definite article *the* accounts for approx. 6% of all words in (British) English, including punctuation and numbers. Verifying this claim on the LOB (the British equivalent of the Brown corpus, see Article 22), we find highly significant evidence against H_0. In particular, there are $f = 68{,}184$ instances of *the* in a sample of $n = 1{,}149{,}864$ words. A two-sided binomial test for $H_0 : \pi = 6\%$ rejects the null hypothesis with a p-value of $p \approx 0.1\%$.

However, the MLE for the true proportion π is actually very close to 6%, viz. $\hat{\pi} = 5.93\%$, and the 95% confidence interval is $\pi = 5.89\% \ldots 5.97\%$. This difference is certainly not of scientific relevance, and $\hat{\pi}$ as well as the entire confidence range would be understood to fall under Baayen's claim of "approximately 6%". The highly significant rejection is merely a consequence of the large sample size and the corresponding high power of the binomial test. Gries (2005) is a recent discussion of the "significance" of statistical significance in corpus work.

At the opposite end of the scale, it is sometimes important to keep the sample size as small as possible, especially when the preparation of the sample involves time-consuming manual data annotation. Power calculations, which are provided by many statistical software packages, can be used to predict the minimum sample size necessary for a reliable rejection of H_0, based on our conjectures about the true effect size.

4 The normal approximation

Looking back at Figure 1, we can see that the binomial sampling distribution has a fairly simple and symmetric shape, somewhat reminiscent of the outline of a bell. The peak of the curve appears to be located at the expected frequency $e = 15$. For other parameter values π and n, we observe the same general shape, only stretched and/or translated. This bell-shaped curve can be described by the following mathematical function:

$$f(x) := \frac{1}{\sigma\sqrt{2\boldsymbol{\pi}}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

This is the formula of a *normal* or *Gaussian distribution* (DeGroot/Schervish 2002, Section 5.6). The parameter μ, called the *mean*, will determine the peak of the bell-shaped curve, and the parameter σ, called the *standard deviation*, will determine the width of the curve (the symbol $\boldsymbol{\pi}$ in this formula stands for Archimedes' constant $\boldsymbol{\pi} = 3.14159\ldots$ and not for a population proportion; to avoid another ambiguity, we write $\exp\bigl(-\frac{(x-\mu)^2}{2\sigma^2}\bigr)$ for the exponential function in lieu of the more commonly encountered $e^{-(x-\mu)^2/2\sigma^2}$, since we are using e to denote the expected frequency). The roles of the two parameters are illustrated in Figure 3.

A binomial distribution with parameters n and π is approximated by a normal distribution with parameters $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1-\pi)}$. Figure 4 shows the same binomial distribution illustrated in Figure 1 (with sample size $n = 100$ and proportion $\pi = 15\%$) and the corresponding normal approximation with parameters $\mu = 15$ and $\sigma \approx 3.57$. The quality of the approximation will increase with sample size and it will depend on π not being too skewed (i.e., not too close to 0 or 1). A rule of thumb might be to trust the approximation only if $\sigma > 3$, which is the case in our example (if you refer back to the formula for σ, you will notice that it depends, indeed, on n and the skewness of π).

The parameters of the normal approximation can be interpreted in an intuitive manner: μ coincides with the expected frequency e under H_0 (remember from Section 2 that e is also given by $n\pi$; we will use μ when referring to the normal distribution formula, e otherwise,

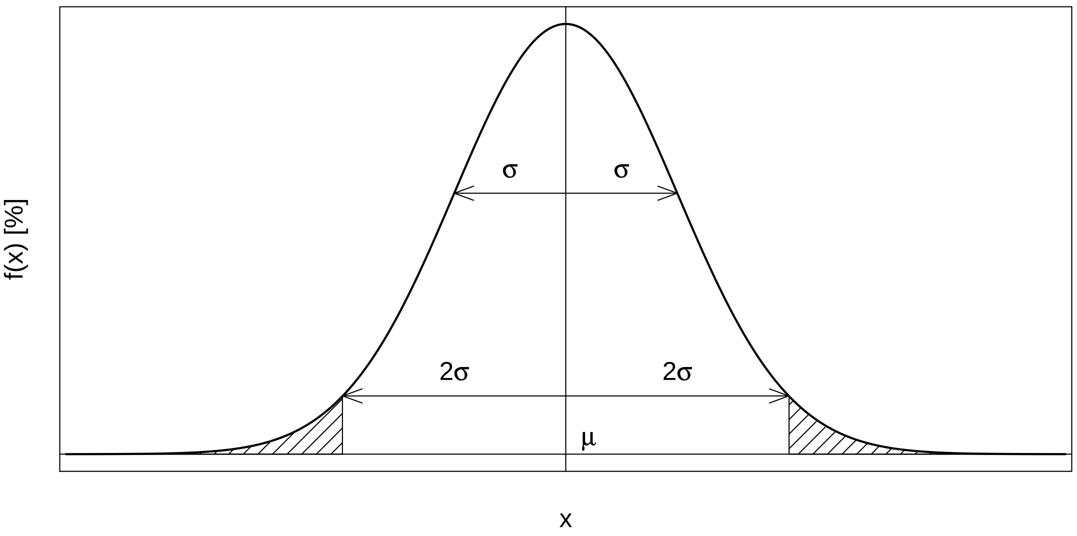

Figure 3: Interpretation of the parameters μ and σ of the normal distribution.

but keep in mind that the two symbols denote the same quantity). σ tells us how much random variation we have to expect between different samples. Most of the samples will lead to observed frequencies between $\mu - 2\sigma$ and $\mu + 2\sigma$ and virtually all observed values will lie between $\mu - 3\sigma$ and $\mu + 3\sigma$ (refer to Figure 3 again), provided that H_0 is true.

To compute binomial tail probabilities based on a normal approximation, one calculates the corresponding area under the bell curve, as illustrated in Figure 4 for the tail probability $\Pr(X \geq 19)$. In this illustration, we have also applied *Yates' continuity correction* (DeGroot/Schervish 2002, Section 5.8), which many statistical software packages use to make adjustments for the discrepancies between the smooth normal curve and the discrete distribution that is approximated. In our example, Yates' correction calculates the area under the normal curve for $x \geq 18.5$ rather than $x \geq 19$.

We find that the normal approximation gives a one-tailed p-value of 16.3% for observed frequency $f = 19$, sample size $n = 100$ and null hypothesis $H_0 : \pi = 15\%$. This is the same p-value we obtained from the (one-tailed) binomial test, indicating that the approximation is very good. Given that the normal distribution (unlike the binomial!) is always symmetrical, the two-tailed p-value can be obtained by simply multiplying the one-tailed value by two (which corresponds to adding up the tail areas under the curve for values that are at least as extreme as the observed value, with respect to e). In our case this gives 32.6%, again equivalent to the binomial test result.

There are two main reasons why the normal approximation is often used in place of the binomial test. First, the exact (non-approximated) binomial test and binomial confidence intervals require computationally expensive procedures that, for large sample sizes such as those often encountered in corpus-based work, can be problematic even for modern computing resources (a particularly difficult case is the extension of confidence intervals to the two-sample setting that we introduce in Section 5 and beyond). Second, the normal approximation leads to a more intuitive interpretation of the difference $f - e$ between observed and expected frequency, and the amount of evidence against H_0 that it provides (the importance of a given raw difference value depends crucially on sample size and on the null hypothesis proportion π_0, which makes it hard to compare across samples and

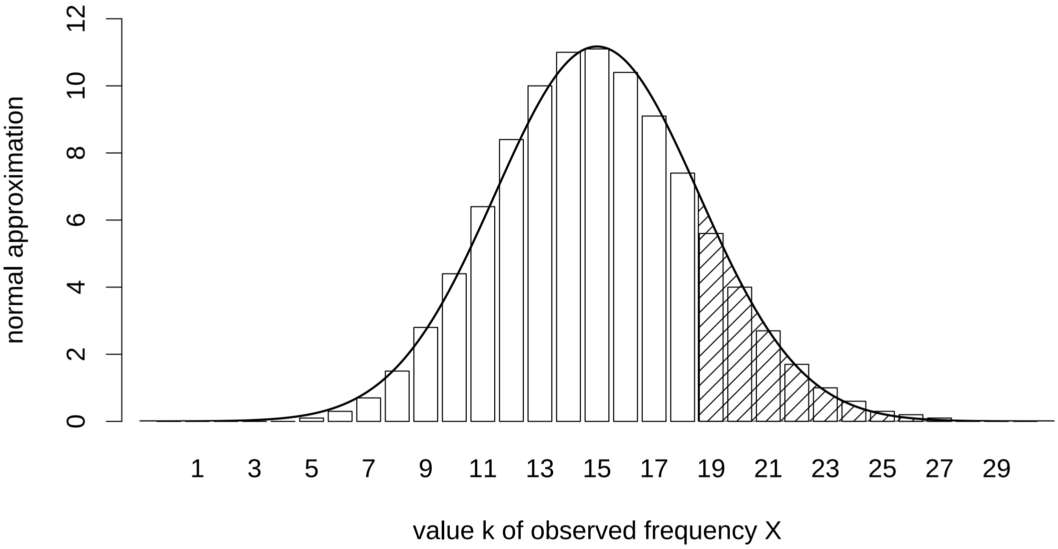

Figure 4: Approximation of binomial sampling distribution by normal distribution curve.

experiments).

An interpretation of $f - e$ (or, equivalently $f - \mu$) that is comparable, e.g., between samples of different sizes, is achieved by a normalized value, the *z-score*, which divides $f - \mu$ by the standard deviation σ (you can think of this as expressing $f - \mu$ in σ's, i.e., using σ as the "unit of measurement"):

$$z := \frac{f - \mu}{\sigma} \tag{5}$$

If two observations f_1 and f_2 (possibly coming from samples of different sizes and compared against different null hypotheses) lead to the same z-score $z_1 = z_2$, they are equally "extreme" in the sense that they provide the same amount of evidence against their respective null hypothesis (as given by the approximate p-values). To get a feel for this, refer back to Figure 3, which illustrates the approximate two-tailed p-value corresponding to $z = 2$ as a shaded area under the normal curve. This area has exactly the same size regardless of the specific shape of the curve implied by H_0 (in the form of the parameters μ and σ). In other words, whenever we observe a value that translates into a z-score of $z = 2$ (according to the respective null hypothesis), we will obtain the same p-value from (the normal approximation to) the binomial test. Since we apply a two-tailed test, an observation that is two standard deviations to the left of the expected value ($z = -2$) will also lead to the same p-value.

Once an observation f has been converted into a z-score z, it is thus easy to decide whether H_0 can be rejected or not, by comparing $|z|$ with previously established *thresholds* for common significance levels α. For $\alpha = .05$, the (two-tailed) z-score threshold is 1.96, so the rejection criterion is $|z| \geq 1.96$; for $\alpha = .01$ the threshold is $|z| \geq 2.58$ and for $\alpha = .001$ it is $|z| \geq 3.29$. Thus, no matter what the original values of f, π and n are, if in an experiment we obtain a z-score of, say, $z = 2$ (meaning that f is two standard deviations away from e), we immediately know that the result is significant at the .05 significance level, but not at the .01 level. Statistics textbooks traditionally provide lists of z-score thresholds corresponding to various significance levels, although nowadays p-values for arbitrary z-scores can quickly be obtained from statistical software packages.

5 Two-sample tests

Until now, we analyzed what is known as a *one-sample* statistical setting, where our null hypothesis concerns a certain quantity (often, the proportion of a certain phenomenon) in the set of all relevant units (e.g., all the sentences of English) and we use a sample of such units to see if the null hypothesis should be rejected. However, *two-sample* settings, where we have two samples (e.g., two corpora with different characteristics, or two sets of sentences of different types) and want to know whether they are significantly different with respect to a certain property, are much more common. Coming back to the example of passivization in idiomatic vs. non-idiomatic constructions from the introduction, our two samples would be sets of idiomatic and non-idiomatic constructions; we would count the number of passives in both sets; and we would verify the null hypothesis that there is no difference between the proportion of passives in the two samples.

It is easier to motivate one aspect of hypothesis testing that is often counter-intuitive, i.e., the fact that we pick as null hypothesis the "uninteresting" hypothesis that we hope to reject, when looking at the two-sample case. First, most linguistic theories, especially categorical ones, are more likely to predict that there is *some* difference between two sets, rather than making quantitative predictions about this difference being of a certain size. Second, in this way, if we can reject the null hypothesis, we can claim that the hypothesis that there is no difference between the groups is not tenable, i.e., that there *is* a difference between the groups, which is what our theory predicts. If, instead, we tested the null hypothesis that there is a certain difference between the groups, and we found that this hypothesis cannot be rejected, we could only claim that, for now, we have not found evidence that would lead us to reject our hypothesis: clearly, a weaker conclusion.

Probably the majority of questions that are of interest to linguists can be framed in terms of a two-sample statistical test: for several examples of applications in syntax, see Article 45; for an application to the study of collocations, see Article 57. Here, we discuss the example of the distribution of passives in two broad classes of written English, "informative" prose (such as daily press) and "imaginative" prose (such as fiction). One plausible *a priori* hypothesis is that these two macro-genres will differ in passive ratios, with a stronger tendency to use passives in informative prose, due to the impersonal, more "objective" tone conferred to events by passive voice (for a more serious corpus-based account of the distribution of the English passive, including register-based variation, see Biber/Johansson/Leech/et al. 1999, Sections 6.4 and 11.3). Our null hypothesis will be that there is no difference between the proportion of passives in informative prose π_1 and imaginative prose π_2, i.e., $H_0 : \pi_1 = \pi_2$. Conveniently, the documents in the Brown corpus are categorized into informative and imaginative writing – thus, we can draw random samples of $n_1 = 100$ sentences from the informative section of the corpus, and $n_2 = 100$ sentences from the imaginative section. Counting the passives, we find that the informative sample contains $f_1 = 23$ passives, whereas the imaginative sample contains $f_2 = 9$ passives.

Since there is a considerable difference between f_1 and f_2, we are tempted to reject H_0. However, before we can do so, we must find out to what extent the difference can be explained by random variation, i.e., we have to calculate how likely it is that the two samples come from populations with the same proportion of passives, as stated by the null hypothesis (statistics textbooks will often phrase the null hypothesis directly as: the samples *are* from the same population). In order to calculate expected frequencies, we have to estimate this common value from the available data, using maximum-likelihood estimation: $\hat{\pi} = (f_1 + f_2)/(n_1 + n_2) = 32/200 = 16\%$ (we sum the f's and n's because, if H_0 is right, then we can treat all the data we have as a larger sample from what, for our purposes, counts as the same population). Replacing H_0 by the more specific null hypothesis $H_0' : \pi_1 = \pi_2 = 16\%$, we can compute the expected frequencies under the null hypothesis, i.e., $e_1 = e_2 = 100 \cdot \hat{\pi} = 16$ (which are identical in our case since $n_1 = n_2 = 100$), as well as the binomial sampling distributions.

In the one-sample case, we looked at the overall probability of f and all other possible values that are more extreme than $|f - e|$. The natural extension to the two-sample case would be to look at the overall probability of the pair (f_1, f_2) and all the other possible pairs of values that, taken together, are more extreme than the sum of $|f_1 - e_1|$ and $|f_2 - e_2|$. The lower this probability, the more confident we can be that the null hypothesis is false. In our case, $|f_1 - e_1|$ and $|f_2 - e_2|$ are directly comparable and might be added up in this way, since the expected frequencies $e_1 = e_2 = 16$ and the sample sizes $n_1 = n_2 = 100$ are the same. However, in many real life situations, we will have to deal with samples of (sometimes vastly) different sizes (e.g., if one of the conditions is relatively rare so that only few examples can be found).

Fortunately, we know a solution to this problem from Section 4: z-scores provide a measure of extremeness that is comparable between samples of different sizes. We thus compute the z-scores $z_1 = (f_1 - e_1)/\sigma_1$ and $z_2 = (f_2 - e_2)/\sigma_2$ (with σ_1 and σ_2 obtained from the estimate $\hat{\pi}$ according to H_0'). For mathematical reasons, the total extremeness is computed by adding up the squared z-scores $x^2 := (z_1)^2 + (z_2)^2$ instead of the absolute values $|z_1|$ and $|z_2|$. It should be clear that the larger this value is, the less likely the null hypothesis of no difference in population proportions is, and thus we should feel more confident in rejecting it. More precisely, the p-value associated with x^2 is the sum over the probabilities of all outcomes for which the corresponding random variable $X^2 := (Z_1)^2 + (Z_2)^2$ is at least as large as the observed x^2, i.e. $\Pr(X^2 \geq x^2)$.

Instead of enumerating all possible pairs of outcomes with this property, we can again make use of the normal approximation, which leads to a so-called *chi-squared distribution* with one *degree of freedom* (df = 1). Using the chi-squared distribution, we can easily calculate the p-value corresponding to the observed x^2, or compare x^2 with known rejection thresholds for different significance levels (e.g. $x^2 \geq 3.84$ for $\alpha = .05$ or $x^2 \geq 6.63$ for $\alpha = .01$). This procedure is known as *(Pearson's) chi-squared test* (Agresti 1996, Section 2.4; DeGroot/Schervish 2002, Sections 9.1-4).

An alternative representation of the observed frequency data that is widely used in statistics takes the form of a so-called *contingency table*:

	sample 1	sample 2
passives	f_1	f_2
other	$n_1 - f_1$	$n_2 - f_2$

(6)

The cells in the first row give the frequencies of passives in the two samples, while the cells in the second row give the frequencies of all other sentence types. Notice that each column of the contingency table adds up to the respective sample size, and that $\hat{\pi}$ (the estimated population proportion under H_0 needed to compute expected frequencies) can be obtained by summing over the first row and dividing by the overall total. Thus, the chi-squared statistic x^2 can easily be calculated from such a table (Agresti 1996, Chapter 2; DeGroot/Schervish 2002, Section 9.3) and most statistical software packages expect frequency data for the chi-squared test in this form. Like in the one-sample case, the normal approximation is only valid if the sample sizes are sufficiently large. The standard rule of thumb for contingency tables is that all *expected* cell frequencies (under H_0') must be

≥ 5 (Agresti 1996, Section 2.4.1). In the usual situation in which $\hat{\pi} < 50\%$, this amounts to $n_1\hat{\pi} \geq 5$ and $n_2\hat{\pi} \geq 5$. Statistical software will usually produce a warning when the normal approximation is likely to be inaccurate.

There is also an *exact* test for contingency tables, similar to the binomial test in the one-sample case. This test is known as *Fisher's exact test* (Agresti 1996, Section 2.6). It is implemented in most statistical software packages, but it is computationally expensive and may be inaccurate for large samples (depending on the specific implementation). Therefore, use of Fisher's test is usually reserved for situations where the samples are too small to allow the normal approximations underlying the chi-squared test (as indicated by the rule of thumb above).

In the current example ($f_1 = 23$ and $f_2 = 9$), the contingency table corresponding to the observed data is

	sample 1	sample 2
passives	23	9
other	77	91

(7)

Using a statistical software package, we obtain $x^2 = 6.29$ for this contingency table, leading to rejection of H_0 at the .05 significance level (but not at the .01 level). The approximate p-value computed from x^2 is $p = 1.22\%$, while Fisher's exact test yields $p = 1.13\%$ (with expected frequencies $n_1\hat{\pi} = n_2\hat{\pi} = 16 \gg 5$, we anticipated a good agreement between the exact and the approximate test). We can thus conclude that there is, indeed, a difference between the proportion of passives in informative vs. imaginative prose. Moreover, the direction of the difference confirms our conjecture that the proportion of passives is higher in informative prose.

A particular advantage of the contingency table notation is that it allows straightforward generalizations of the two-sample frequency comparison. One extension is the comparison of more than two samples representing different conditions (leading to a contingency table with $k > 2$ columns). For instance, we might want to compare the frequency of passives in samples from the six subtypes of imaginative prose in the Brown corpus (general fiction, mystery, science fiction, etc.) The null hypothesis for such a test is that the proportion of passives is the same for all six subtypes, i.e. $H_0 : \pi_1 = \pi_2 = \cdots = \pi_6$. Another extension leads to contingency tables with $m > 2$ rows. In our example, we have distinguished between passive sentences on one hand and all other types of sentences on the other. However, this second group is less homogeneous so that further distinctions may be justified, e.g., at least between sentences with intransitive and transitive constructions. From such a three-way classification, we would obtain three frequencies $f^{(p)}$, $f^{(i)}$ and $f^{(t)}$ for each sample, which add up to the sample size n. These frequencies can naturally be collected in a contingency table with three rows. The null hypothesis would now stipulate that the proportions of passives, intransitive and transitives are the same under both conditions (assuming $k = 2$), viz. $\pi_1^{(p)} = \pi_2^{(p)}$, $\pi_1^{(i)} = \pi_2^{(i)}$ and $\pi_1^{(t)} = \pi_2^{(t)}$. In general, an x^2 value can be calculated for any $m \times k$ contingency table in analogy to the 2×2 case. The p-value corresponding to x^2 can be obtained from a chi-squared distribution with $\mathrm{df} = (m-1)(k-1)$ degrees of freedom. If the expected frequency in at least one of the cells is less than 5, a version of Fisher's exact test can be used (this version is considerably *more* expensive than Fisher's test for 2×2 tables, though).

Having established that the proportion of passives is different in informative vs. imaginative prose, we would again like to know how large the effect size is, i.e. by how much the proportions π_1 and π_2 differ. This is particularly important for large samples, where small (and hence linguistically irrelevant) effect sizes can easily lead to rejection of H_0 (cf. the discussion in Section 3). A straightforward and intuitive measure of effect size is the difference $\delta := \pi_1 - \pi_2$. When the sample sizes are sufficiently large, normal approximations can be used to compute a confidence interval for δ. This procedure is often referred to as a *proportions test* and it is illustrated, for example, by Agresti (1996, Section 2.2). In our example, the 95% confidence interval is $\delta = 3.0\% \ldots 25.0\%$, showing that the proportion of passives is at least 3 percentage points higher in informative prose than in imaginative prose (with 95% certainty).

In other situations, especially when π_1 and π_2 are on different orders of magnitude, other measures of effect size, such as the ratio π_1/π_2 (known as *relative risk*) may be more appropriate. A related measure, the *odds ratio* θ, figures prominently because an exact confidence interval for θ can be obtained from Fisher's test. Most software packages that implement Fisher's test will also offer calculation of this confidence interval. In many linguistic applications (where π_1 and π_2 are relatively small), θ can simply be interpreted as an approximation to the ratio of proportions (relative risk), i.e., $\theta \approx \pi_1/\pi_2$. On these measures see, again, Section 2.2. of Agresti (1996). Effect size in general $m \times k$ contingency tables is much more difficult to define, and it is most often discussed in the setting of so-called *generalized linear models* (Agresti 1996, Chapter 4).

Examples of fully worked two-sample analyses based on contingency tables can be found in Articles 45 and 57. As illustrated by Article 45 in particular, contingency tables and related two-sample tests can be tuned to a number of linguistic questions by looking at different kinds of linguistic populations. For example, if we wanted to study the distribution of by-phrases in passive sentences containing two classes of verbs (say, verbs with an agent vs. experiencer external argument), we could define our two populations as all passive sentences with verbs of class 1 and all passive sentences with verbs of class 2. We would then sample passive sentences of these two types, and count the number of by-phrases in them. As a further example, we might be interested in comparing alternative morphological and syntactic means to express the same meaning. For example, we might be interested, with Rainer (2003), in whether various classes of Italian adjectives are more likely to be intensified by the suffix *-issimo* or by the adverb *molto*. This leads naturally to a contingency table for intensified adjectives with *-issimo* and *molto* columns, and as many rows as the adjective classes we are considering (or vice versa). The key to the successful application of statistical techniques to linguistic problems lies in being able to frame interesting linguistic questions in operational terms that lead to meaningful significance testing. The following section will discuss different ways to perform this operationalization.

6 Linguistic units and populations

As we just said, from the point of view of linguists interested in analyzing their data statistically, the most important issue is how to frame the problem at hand so that it can be operationalized in terms suitable for a statistical test. In this Section, we introduce some concepts that might be useful when thinking of linguistic questions in a statistical way.

In the example used throughout the preceding Sections, we have defined the population as the set of all (written American) English sentences and considered random samples of sentences from this population. However, statistical inference can equally well be based on any other linguistic unit, such as words, phrases, paragraphs, documents, etc. This *unit of measurement* if often called a *token* in corpus linguistics, at least when referring to words. Here, we use the term more in general to refer to any unit of interest.

The *population* then consists of all the utterances that have ever been produced (or could be produced) in the relevant (sub)language, broken down into tokens of the chosen type. We might also decide to focus on tokens that satisfy one or more other criteria and narrow down the population to include only these tokens. For instance, we might be concerned with the population of words that belong to a specific syntactic category; or with sentences that contain a particular verb or construction, etc.

What we are interested in is the *proportion* π of tokens (in the population) that have a certain additional property: e.g., word tokens that are nouns, verb tokens that belong to the inflectional paradigm of *to make*, sentences in the passive voice, etc. The properties used to categorize tokens for this purpose are referred to as *types* (in contrast to tokens, which are the categorized objects).

Since the full population is inaccessible, our conclusions have to be based on a *(random) sample* of tokens from the population. Such a sample of language data is usually called a *corpus* (or can be derived from a corpus: when we define the population as a set of verb tokens, for example, our sample might comprise all instances of verbs found in the corpus). The *sample size* n is the total number of tokens in the sample, and the number of tokens that exhibit the property of interest (i.e., that belong to the relevant type) is the observed *frequency* f.

The same observed frequency can have different interpretations (with respect to the corresponding population proportion) depending on the units of measurement chosen as tokens, and the related target population. For instance, the number of passives in a sample could be seen relative to the number of sentences (π = proportion of sentences in the passive voice), relative to the number of verb phrases (π = proportion of passive verb phrases), relative to word tokens (π = relative frequency of passive verb phrases per 1,000 words), relative to all sentences containing transitive verbs (π= relative frequency of actual passives among sentences that could in principle be in passive voice). Note that each of these interpretations casts a different light on the observed frequency data. It is the linguist's task to decide which interpretation is the most meaningful, and to draw conclusions about the linguistic research questions that motivated the corpus study.

Other examples might include counting the number of deverbal nouns in a sample from the population of all nouns in a language; counting the number of words ending in a coronal stop in a sample from the population of all words in the language; counting the number of sentences with heavy-NP shift in a sample from the population of all sentences with a complement that could in principle undergo the process; counting the number of texts written in first person in a sample from the population of literary texts in a certain language and from a certain period. Related problems can also be framed in terms of looking at *two* samples from distinct populations (cf. Section 5), e.g., counting and comparing the number of deverbal nouns in samples from the populations of abstract and concrete nouns; counting the number of words ending in a coronal stop in samples from the population of all native words in the language and the population of loanwords; counting the number of texts written in first person in samples from populations of texts belonging to two different literary genres.

In many cases, frequencies are computed not only for a single property, but for a set of mutually exclusive properties, i.e., a *classification* of the tokens into different types. In the two-sample setting this leads naturally to a $m \times 2$ contingency table (with the types in the classification as rows, and the two populations we are comparing as columns). Note that the classification has to be *complete*, so that the columns of the table add up to the respective sample sizes, which is often achieved by introducing a category labeled "other" (the single-property/two-samples cases above correspond to 2×2 contingency tables with an "other" class: e.g., deverbal vs. "other" nouns compared across the populations of abstract vs. concrete nouns).

As an example of a classification into multiple categories, word tokens might be classified into syntactic categories such as noun, verb, adjective, adverb, etc., with an "other" class for minor syntactic categories and problematic tokens. A chi-squared test might then be performed to compare the frequencies of these categories in samples from two genres. As another example, one might classify sentences according to the semantic class of their subject, and then compare the frequency of these semantic classes in samples of the populations of sentences headed by true intransitive vs. unaccusative verbs. It is not always obvious which characteristics should be operationalized as a classification of the tokens into types, and which should rather be operationalized in terms of different populations the tokens belong to. In some cases, it might make more sense to frame the task we just discussed in terms of the distribution of verb types across populations of sentences with different kinds of subjects, rather than vice versa. This decision, again, will depend on the linguistic question we want to answer.

In corpus linguistics, *lexical classifications* also play an important role. In this case, types are the distinct word forms or lemmas found in a corpus (or sequences of word forms or lemmas). Lexical classifications may lead to extremely small proportions π (sometimes measured in occurrences per million words) and huge differences between populations in the two-sample setting. Article 57 discusses some of the relevant methodologies in the context of collocation extraction.

The examples we just discussed give an idea of the range of linguistic problems that can be studied using the simple methods based on count data described in this Article. Other problems (or the same problems viewed from a different angle) might require other techniques, such as those mentioned in the next two Sections. For example, our study of passives could proceed with a *logistic regression* (see Section 8), where we look at which factors have a significant effect on whether a sentence is in the passive voice or not. In any case, it will be fundamental for linguists interested in statistical methods to frame their questions in terms of populations, samples, types and tokens.

7 Non-randomness and the unit of sampling

So far, we have always made the (often tacit) assumption that the observed data (i.e., the corpus) are a random sample of tokens of the relevant kind (e.g., in our running example of passives, a sentence) from the population. Most obviously, we have compared a corpus study to drawing balls from an urn in Section 2, which allowed us to predict the sampling distribution of observed frequencies. However, a realistic corpus will rarely be built by sampling individual tokens, but rather as a collection of contiguous stretches of text or even entire documents (such as books, newspaper editions, etc.). For example, the Brown corpus consists of 2,000-word excerpts from 500 different books (we will refer to these excerpts as "texts" in the following). The discrepancy between the *unit of measurement* (a token) and the *unit of sampling* (which will often contain hundreds or thousands of tokens) is particularly obvious for lexical phenomena, where tokens correspond to single words. Imagine the cost of building the Brown corpus by sampling a single word each from a million different books rather than 2,000 words each from only 500 different books!

Even in our example, where each token corresponds to an entire sentence, the unit of sampling is much larger than the unit of measurement: each text in the Brown contains roughly between 50 and 200 sentences. This need not be a problem for the statistical analysis, as long as each text is itself a random sample of tokens from the population, or at least sufficiently similar to one. However, various factors, such as the personal style of an author or minor differences in register or conventions within a particular subdomain, may have a *systematic* influence on how often passives are used in different texts. This means that the variability of the frequency of passives between texts may be much larger than between random samples of the same sizes (where all variation is purely due to chance effects).

Again, the problem is most obvious for lexical frequencies. Many content words (except for the most general and frequent ones) will almost only be found in texts that deal with suitable topics (think of nouns like *football* or *sushi*, or adjectives like *coronal*). On the other hand, such topical words tend to have multiple occurrences in the same text, even if these would be extremely unlikely in a random sample (indeed, the "burstiness" of words in specific texts is used as strategy to find interesting keywords; see, e.g., Church 2000).

The increased variability of frequency between the individual texts is attenuated to some extent when corpus frequencies are obtained by summing over all the (randomly selected) texts in a corpus. However, in most cases the corpus frequencies will still show more variation than predicted by the binomial distribution.

In order to verify empirically whether a linguistic phenomenon such as the frequency of passives is subject to such non-randomness, we can compare the distribution of observed frequencies across the texts in a corpus with the distribution predicted for random samples by the binomial distribution. An example of such a comparison is shown in Figure 5. From each Brown text, we have taken a sample of 50 sentences (this subsampling step was necessary because the number of sentences per text varies from around 50 to more than 200). By tabulating the observed frequencies, we obtain the distribution shown as black bars in Figure 5. The gray bars show the binomial distribution that we would have obtained for random samples from the full population (the population proportion of passives was estimated at $\pi = 27.5\%$, based on the Brown data). Note that we have used only informative prose texts, since we already know from Section 5 that the proportion of passives differs considerably between the two major sections of the corpus.

As the figure shows, the observed amount of variation is larger than the one predicted from the binomial distribution: look for example at the proportion of observed and binomial samples with $X \leq 7$. The standard deviation (which, as discussed in Section 4, is a measure of the width of a distribution) is $\sigma = 6.63$ for the empirical distribution, but only $\sigma = 3.16$ for the binomial distribution. The corresponding z-scores, having σ in the denominator (see Equation (5) in Section 4), will be smaller for the empirical distribution, and thus the results are less significant than they would seem according to the binomial distribution. This means that the binomial test will lead to rejection of a true null hypothesis more easily than should be the case, given the spread of the actual distribution.

Suppose that we want to test the null hypothesis $H_0 : \pi = 27.5\%$ (which is in fact true) based on a sample of $n = 50$ sentences from the informative prose in the Brown corpus. If the observed frequency of passives in this sample is $f = 7$, we feel confident to reject H_0 ($e = 13.75$ leads to a z-score of $z = -2.14$, above the $\alpha = .05$ threshold of $|z| \geq 1.96$). However, if all sentences in this sample came from the same text (rather than being sampled randomly from the entire informative prose section), Figure 5 shows that the risk of obtaining *precisely* $f = 7$ by chance is already around 4%! The "true" z-score (based on the standard deviation computed from the observed samples) is only $z = -1.02$, far away from any rejection threshold (in fact, this z-score indicates a risk of more than 30% that H_0 would be wrongly rejected).

Seeing how non-randomness effects can lead to a drastic misinterpretation of the observed frequency data, a question arises naturally: How can we make sure that a corpus study is not affected by non-randomness? While for many practical purposes it might be

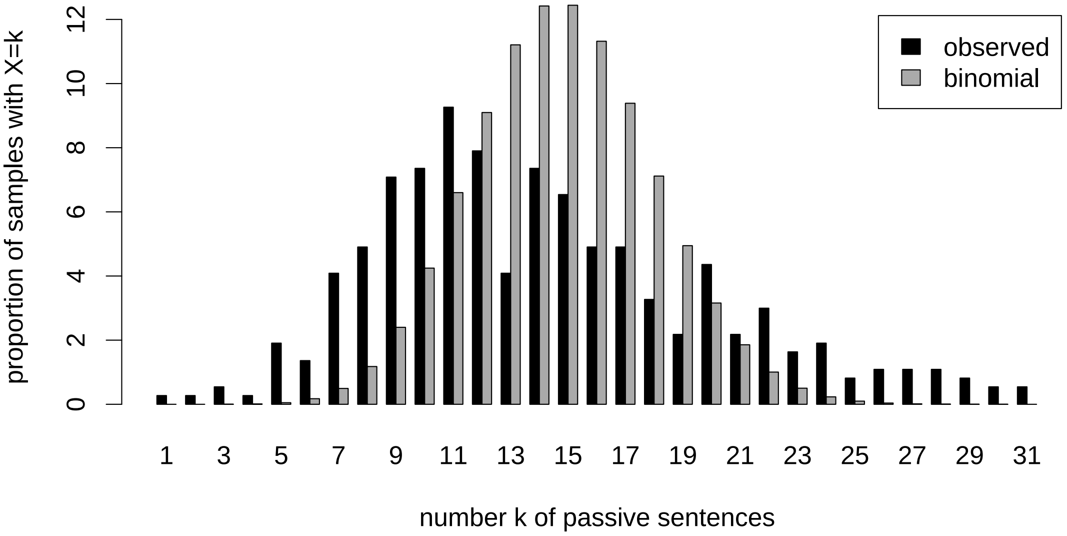

Figure 5: Comparison of the frequencies of passives in the texts of the Brown corpus (informative prose only) with the binomial distribution predicted for random samples. In order to ensure comparability of the frequencies, 50 sentences were sampled from each Brown text.

possible to ignore the issue, the only way to be absolutely sure is to ascertain that the unit of sampling coincides with the unit of measurement. When using a pre-compiled corpus (as will be the case for most studies in corpus linguistics) or when it would be prohibitively difficult and time consuming to sample individual tokens, we have no choice but to adjust the *unit of measurement*. For example, when our data are based on the Brown corpus, the unit of sampling – and hence the unit of measurement – would be a text, i.e., a 2,000-word excerpt from a coherent document. Of course, we can no longer classify such an excerpt as "passive" or "non-passive". Instead, what we observe for each token is a real-valued number: the proportion of passive sentences in the text.

Unlike previously, where each measurement was essentially a yes/no-decision ("passive" or "not passive") or a m-way classification, measurements are now real-valued numbers that can in principle assume any value between 0 and 1 (3/407, 139/211, etc.). Statisticians speak of a *nominal scale* (for yes/no-decisions and classifications) vs. an *interval scale* (for numerical data). In order to analyze such data, we need an entirely different arsenal of statistical procedures, such as the well-known *t-test*. These methods are explained in any introductory textbook on statistics, and we give a brief overview of the most important terms and techniques in Section 8.

This approach is only viable for phenomena, such as passive voice, that have a reasonably large number of occurrences in each text. It would not be sensible to count the proportion of occurrences of the collocation *strong tea* in the Brown texts (or even in a corpus made of larger text stretches), since the vast majority of texts would yield a proportion of 0% (in the Brown corpus, *strong tea* occurs exactly once, which means that in all texts but one the proportion will indeed be 0%).

Notice that, from a statistical perspective, the issues of representativeness and balance sometimes discussed in connection to corpus design (see Article 11) involve two aspects: 1) How to define the target population precisely (is it possible to delimit a set of utterances that constitutes the population of, say, "contemporary English"?), and 2) how to take a random sample from the target population (with the complication discussed in this section that what might constitute a random sample of, say, documents, will not be a random sample of, say, sentences). See Evert (2006) for an extended discussion of non-randomness in corpus linguistics.

8 Other techniques

Like for count data, there is a range of statistical tests that can be used to analyze data on an interval scale (such as the relative number of sentences containing passives per document discussed in the previous Section, or reaction times from a psycholinguistic experiment). For the one-sample case, in which we want to test whether an observed interval-scale quantity (such as the proportion of passive sentences in a text) could plausibly come from a population where the same quantity has a certain distribution with a specific mean and standard deviation, you can use a *(one sample) t-test* (comparable to the binomial test for count data or, more precisely, the normal approximation based on z-scores). Unsurprisingly, when two samples are compared, the appropriate test is a *two-sample t-test* (corresponding to the chi-squared test for count data). However, in order to compare more than two samples, rather than performing a series of pairwise t-tests (a procedure that would make it much more likely that we obtain a significant result by chance), the technique to be applied is the *one-way analysis of variance (ANOVA).* The ANOVA can only tell us whether at least one sample in the set is different from at least one other sample, and post-hoc tests must then be performed to identify the sample(s) responsible for rejection of H_0.

In some settings, the variables in two samples have a natural pairing. For example, if we compare the proportion of passives in English and Spanish texts based on a parallel corpus, we should make use of the information that the texts are paired, so that we control for irrelevant factors that may affect passive proportion (e.g., style and topic of a text) which should have a similar weight in an original and its translation. The appropriate test, in this case, is the *paired t-test.*

In many studies, it makes sense to operationalize the problem as one of assessing the association between two properties of the same unit, both measured on an interval scale. For example, we might be interested in the issue of whether there is a relation between the proportion of passives and, say, that of nominalizations (as both are plausible markers of more formal registers). Given a list of pairs specifying the (relative) frequencies of passives and nominalizations in each of the texts in our sample, we can perform a *correlation analysis.* In this case, the null hypothesis will be that there is no correlation between the two properties; and effect size will be measured in terms of how much of the variability of one variable can be explained by linear dependence on the other variable (standard correlation analysis will not capture *nonlinear* relations between variables).

A significant correlation does not imply a causal relation between two variables (even if the numbers of passives and nominalizations turn out to be correlated, it is unlikely that passives "cause" nominalization or vice versa). Often, however, we want to go beyond the mere observation of a relationship between two variables. If we hypothesize that the behavior of a certain variable depends on that of one or more other variables, we will want to use statistics to test whether our *independent* variables predict the values of the *dependent* variable beyond chance level. In this case, we use the technique of *(multiple) linear regression* (which is related to correlation). In linear regression, the independent variables can be a mixture of discrete and continuous variables, but the dependent variable must be continuous.

Similar techniques can also be applied to the analysis of the kind of categorical data (resulting in a contingency table of frequency counts) that have been the focus of this Article. The equivalent of linear regression in this case is *logistic regression.* For example, a logistic regression analysis could try to predict whether a sentence is in passive voice or not (a dichotomous dependent variable) in terms of factors such as the semantic class of the verb (a categorical variable), the overall entropy of the sentence (a continuous variable), etc. A full regression analysis tests significant effects of the independent variables, but typically it also checks that the independent variables are not correlated with each other, and it might look for the optimal combination of independent variables.

The cases we listed here (detection of differences and estimation of population values in one/two/multiple paired/non-paired sample cases, assessment of association/correlation between variables, regression) constitute a nearly exhaustive survey of the analytical settings considered in inferential statistics. More advanced techniques, rather than introducing completely new scenarios, will typically deal with cases in which one or more of the assumptions of the basic models are not met or the models need to be extended. For example, more sophisticated ANOVA models can take multiple categorizations of the data and their interactions into account (akin to the analysis of $m \times k$ contingency tables with m and/or k greater than 2). Advanced regression techniques can detect non-linear relations between the dependent and independent variables. So-called "distribution-free" tests make no assumption about the distribution of the underlying population(s) nor about the sampling distribution (these are typically referred to as *non-parametric methods*). Simulation-based methods (*Monte Carlo methods*, the *Bootstrap*) provide an alternative to analytical estimation of various parameters. A wealth of exploratory and visual methods are available to evaluate the validity of assumptions and the quality of the resulting models. *Bayesian inference*, a very important branch of statistics, allows, among other things, to distinguish between "more plausible" and "less plausible" estimates within a confidence interval (the classic binomial confidence interval described in Section 3 indicates a range of plausible values for the population proportion, but it does not distinguish among these values, whereas, intuitively, we would consider the MLE proportion much more plausible than, say, the values at the edges of the confidence interval).

Some important kinds of corpus data, such as distributions of word types, are characterized by the presence of a very large number of very rare types (words that occur only once or never at all in the corpus at hand) and few extremely frequent types (function words). These extremely skewed distributions make the application of standard statistical models to certain tasks problematic (mainly, estimating the number of word types in a population as well as related quantities), and demand specialized statistical tools. For a general survey of the problems involved, see Article 39 and the references on statistical modeling of word frequency distributions recommended there.

Almost every elementary statistics textbook (including those listed in the next Section) will introduce t-tests, ANOVA, correlation and regression. Advanced techniques are nowadays within easy reach of non-statisticians thanks to their implementation in user-friendly software packages. Here, we would like to stress once more that, for all of the large variety of available procedures and their complications, the basic logic of hypothesis testing and estimation is essentially the same that we illustrated with very simple examples of frequency count data in the first Sections of this Article. It is not essential to know mathematical details of all the techniques in order to apply them, but it is important to understand the basic principles of hypothesis testing and estimation; the assumptions of a test, its null hypothesis, and the meaning of a p-value; and to make sure that the assumptions are met by the data and that the research question can be translated into a meaningful null hypothesis. And, of course, the linguistic interpretation of the statistical results is at least as crucial as the correctness of the methods applied.

We have focused here on statistical inference for hypothesis testing and estimation, as applied to corpus data. This is only a part, albeit a fundamental one, of the role that statistical methods play in corpus-related disciplines today. For a survey of statistical procedures used for *exploratory* purposes (i.e., as an aid in uncovering interesting patterns in the data), see Articles 40 and 42. Statistical methods also play a very important role as *modeling* tools for machine learning techniques applied to natural language (Article 41) and more generally in so-called empirical natural language processing (see, e.g., Article 50 on machine translation, and Manning/Schütze 1999 for an introduction to statistical NLP).

9 Directions for further study

A much more in-depth introduction to the statistical inference methods appropriate for count data that we discussed here is provided by Agresti (1996) or, at a more technical level, Agresti (2002). There is, of course, a vast number of introductory general statistics books. DeGroot/Schervish (2002) present a particularly thorough and clear introduction, although it requires at least a basic mathematical background. Among the less technical introductions, we recommend the one by Hays (1994), a book that provides non-mathematical but rigorous explanations of the most important notions of statistical inference (although it focuses on the statistical methods for the analysis of experimental results, which are only partially relevant to corpus work). There is also a wealth of statistics "cookbooks" that illustrate when to apply a certain technique, how to apply it, and how to interpret the results. These are often and usefully linked to a specific statistical software package. For example, Dalgaard (2002) is an introduction to running various standard statistical procedures in R (see below).

There are a few older introductions to statistics explicitly geared towards linguists. The one by Woods/Fletcher/Hughes (1986) is a classic, whereas the one by Butler (1985) has the advantage of being now freely available on the Web:

`http://www.uwe.ac.uk/hlss/llas/statistics-in-linguistics/bkindex.shtml`

Older introductions tend to focus on techniques that are more relevant to psycholinguistics, phonetics and language testing than to corpus analysis. Oakes (1998) presents a survey of applications of statistics in corpus studies that trades depth for wider breadth of surveyed applications and methods. It is likely that, with the growing interest in corpora and statistical approaches to linguistics in general, the next few years will see the appearance of more statistics textbooks targeting corpus linguists.

There is nowadays a large number of statistical software packages to choose from. We recommend R:

`http://www.r-project.org/`

R supports an impressive range of statistical procedures and, being open-source and available free of charge, it is attracting a growing community of developers who add new functionalities, including some that are of interest to corpus linguists. These extensions range from advanced data visualization techniques to modules explicitly targeting corpus work, such as the `corpora` library developed by the authors of this Article:

`http://purl.org/stefan.evert/SIGIL`

The `corpora` library (also free and open-source) provides support for carrying out the statistical analyses described in this Article (the Web site has a tutorial that shows how to run them), as well as several sample data sets. There is an increasing number of introductory textbooks with concrete R examples, and we know of several R-based books focusing on statistical methods in linguistics that are currently in preparation. Shravan Vasishth has written (and is constantly updating) an online book aimed at (psycho-)linguists, that introduces statistics in R through a simulation approach. This book is freely available (under a Creative Commons license) from:

`http://www.ling.uni-potsdam.de/~vasishth/SFLS.html`

Finally, Wulff (2005) provides a survey of online statistics facilities.

References

Agresti, Alan (1996), *An Introduction to Categorical Data Analysis.* Chichester: Wiley.

Agresti, Alan (2002), *Categorical Data Analysis, second edition.* Chichester: Wiley.

Baayen, Harald (2001), *Word Frequency Distributions.* Dordrecht: Kluwer.

Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus Linguistics.* Cambridge: Cambridge University Press.

Biber, Douglas/Johansson, Stig/Leech, Geoffrey/Conrad, Susan/Finegan, Edward (1999), *Longman Grammar of Spoken and Written English.* Harlow, UK: Pearson Education.

Butler, Christopher (1985), *Statistics in Linguistics.* Oxford: Blackwell.

Chomsky, Noam (1986), *Knowledge of Language: Its Nature, Origins, and Use.* New York: Praeger.

Church, Kenneth (2000), Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p^2. In *Proceedings of the 17th Conference on Computational Linguistics*, 180-186.

Culicover, Peter/Jackendoff, Ray (2005), *Simpler Syntax.* Oxford: Oxford University Press.

Dalgaard, Peter (2002), *Introductory Statistics with R.* New York: Springer.

DeGroot, Morris/Schervish, Mark (2002), *Probability and Statistics, third edition.* Boston: Addison-Wesley.

Evert, Stefan (2006). How random is a corpus? The library metaphor. In: *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 177-190.

Hays, William (1994), *Statistics, fifth edition.* New York: Harcourt Brace.

Gries, Stephan (2005), Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. In: *Corpus Linguistics and Linguistic Theory*, 1, 277-294.

Manning, Christopher/Schütze, Hinrich (1999), *Foundations of Statistical Natural Language Processing.* Cambridge (Mass.): MIT Press.

McEnery, Tony/Wilson, Andrew (2001), *Corpus Linguistics, second edition.* Edinburgh: Edinburgh University Press.

Oakes, Michael (1998), *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Pearson, Egon (1990), *'Student': A Statistical Biography of William Sealy Gosset.* Oxford: Clarendon Press.

Rainer, Franz (2003), Studying restrictions on patterns of word-formation by means of the Internet. In: *Rivista di Linguistica*, 15, 131-139.

Schütze, Carson (1996), *The Empirical Base of linguistics: Grammaticality Judgments and Linguistic Methodology.* Chicago: University of Chicago Press.

Woods, Anthony/Fletcher, Paul/Hughes, Arthur (1986), *Statistics in Language Studies.* Cambridge: CUP.

Wulff, Stefanie (2005), Online statistics labs. In: *Corpus Linguistics and Linguistic Theory*, 1, 303-308.