Chapter 5. Predictions from data

By "prediction" in the context of this book is meant the estimation, based on samples, of the richness and other parameters for communities. It also means, in the context of the first section, the estimation of a sample parameter from other parameters as a check on the accuracy of the logistic-J distribution as a descriptor of samples.

In the first section I show how the logistic-J distribution predicts the maximum abundance accurately (in the statistical sense) on the basis of other parameters, such as the mean abundance and the height of the initial peak (at lowest abundance). In subsequent sections I take a closer look at the sampling process, explaining why no richness estimation process can succeed without knowing both the sample intensity and the underlying distribution. I then illustrate this requirement with a brief review of some richness estimation methods already in use.

The second half of the chapter describes and compares two procedures for estimating community richness on the basis of sample richness. Computer experiments make it possible not only to determine the relative effectiveness of the procedures, but to determine the contributions to sample variance from irregularities present in the community being sampled as well as those contributed by the sampling process itself. Distinguishing the two sources of variation, as well as their influence on the sample, is a crucial step on the way to a fuller understanding of the sampling process in its entirety.

Ideally, predictions from theory, at least numerical ones, should come equipped with error bounds so that biologists who make such predictions have a reliable estimate of the uncertainty that accompanies them. Ecologists have been slow in coming to the realization enunciated by Doak, Estes et al. (2008): "Over the last decade, there has been increasing recognition that ecological predictions must be advanced with clear statements of their uncertainty." Such error bounds are a strict requirement of what I have called exact ecology. In Sections 5.4 and 5.5 it is shown how to derive error estimates based on interval statistics arising from samples.

5.1 predicting maximum abundance

The logistic-J distribution has two parameters that, in the context of the metastudy (see Chapter 7), were estimated on the basis of the mean μ of a sample and its initial peak F_1. Neither of these measurements has any obvious relationship with the maximum abundance Δ. In the absence of some theoretical relationship, one would be hard put to make any prediction whatever about the size of Δ on the basis of either of these measurements, or both, for that matter. Yet, when the transfer equations (See Appendix B.10.) are solved with the mean μ and F_1 as input, the values of ε and Δ that correspond to F_1 and μ in the logistic-J function are produced. The parameter ε is closely related to F_1 via R, but not readily discernible in a sample histogram. The parameter Δ, on the other hand, is directly visible, in a sense, as the largest abundance Δ' in the sample; the value of Δ' will be a good estimator of Δ if the logistic-J theory is correct, so it forms a relevant

test for the theory. How well does Δ' predict Δ? Very well indeed, if the ratio Δ'/Δ, expressed as a percentage, equals 100.0%.

Over a great many samples, how close does Δ get to Δ', on average? The metastudy explained in Chapter 7 incorporated the sub-experiment of recording the ratios Δ'/Δ, expressed as percentages, over all 125 biosurveys. If logistic-J theory is correct, it should predict maximum abundances accurately in the sense above.

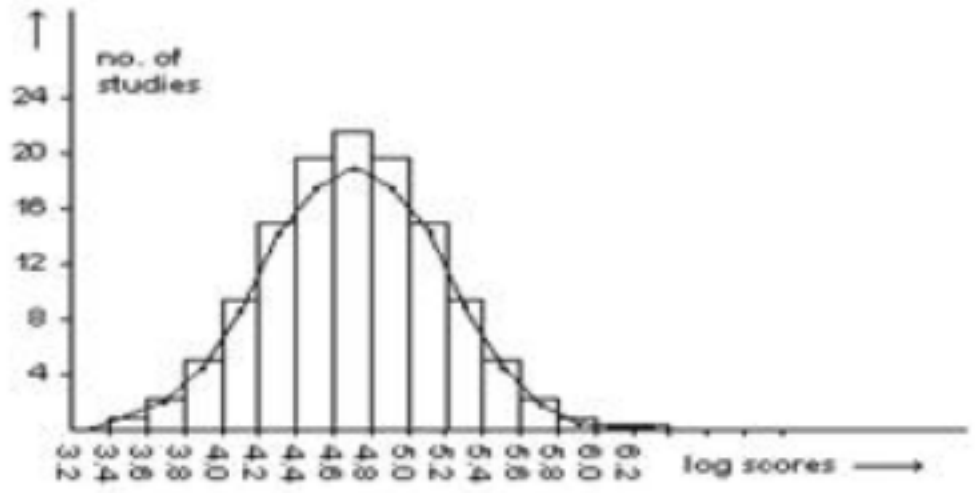

Figure 5.1. Distribution of sizes of delta ratios

Figure 5.1 shows a histogram of predictions of the logistic-J distributions for the 125 biosurveys. Each prediction consists of the ratio of actual abundance Δ' to the predicted abundance Δ in the form of a percentage. However, because ratios are essentially multiplications, I used the lognormal distribution as the appropriate envelope in this context. (Limpert et al. 2001) This meant taking the natural logarithm of the score ratios and plotting them, as in Figure 5.1. Superimposed on this plot is the normal distribution N[4.5, 0.46], indicated by the polygonal arc. The standard deviation of 0.46 was calculated from the data. The logarithm of the actual mean is 4.64 but, when fitted against the empirical data, appeared to be too far to the right, with all the bars to the left of the mean well above their corresponding theoretical values, while those to the right appeared too low. The main purpose of the plotted values, however, is to demonstrate that the data has the right overall shape for this treatment and therefore reflects a "natural' outcome.

When subjected to a chi square test, the empirical vs theoretical data yielded an overall score of 5.966, somewhat below the expected figure of 7.0 for a test with seven degrees of freedom and well below the 95% rejection level of 14.067.

The average prediction for Δ in the collection of 125 biosurveys turns out to be 101.9, well

within the expected range of possibilities for a correct theory but very close to the most likely value (100.0) in relation to the standard deviation of the ratios Given the high vriance of the percentsage ratios, (sd = 53.6) the closeness of this result is almost accidental.

While the foregoing result is entirely consistent with the prediction of logistic-J theory of Δ as the average maximum abundance, it must be remarked that the actual abundance occurs more frequently in the range lower than 100.0 than above it. This is due to the fact that, while the size of ratios less than 100.0 is limited, the size above it is not, as witness the highest value of 426.6 in the set of 126 percentage ratios.

5.2 Predicting species richness

In the remainder of this chapter, the reader will encounter two epsilons and two deltas, those of the community and of the sample. The community parameters will be indicated by ε and Δ, while those of the sample will be indicated by ε' and Δ', respectively.

Suppose that a community C of N organisms inhabits an area A and that a biologist wishes to know the number R of species in C, based on samples of it. As will be seen shortly, he or she must know both the intensity r of the sample, as well as the distribution underlying the community in order to have any hope of a realistic estimate for the number R.

I will present counterexamples to the possibility that such knowledge is not necessary to the project of determining R. The relative ease with which these counterexamples are produced is a strong indication that the project is not even approximately practical.

Example 1

The first counterexample shows that knowing the distribution underlying the community is critical to the determination of R. It involves two rather different artificial communities, C_1 and C_2, that have the same number N of individuals, but with greatly differing numbers of species. Yet, when sampled, C_1 and C_2 yield the same number of species, at least most of the time.

Let C_1 have the univoltine distribution with 11 species, all having abundance 10, and let C_2 be the uniform community with 20 species, two in each abundance category, from 1 to 10. One distribution is a vertical spike, the other is long and perfectly flat. It is easy to verify that N = 110 in both cases so that, when sampled at intensity r = 0.1, the sample size is 11 in both cases.

Figure 5.2 displays histograms of the two communities. Because of the small size of these communities, the sample will be made with replacement. The terminology will reflect this, with the term "draw" being replaced by "choose".

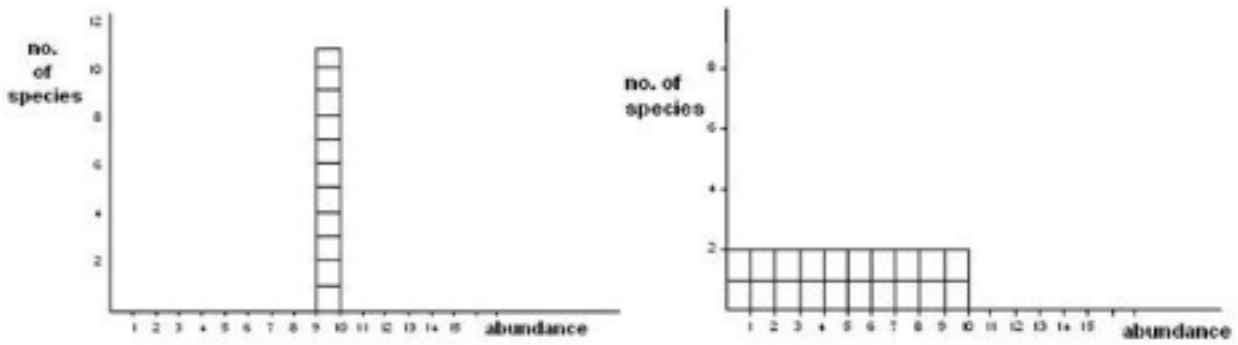

Figure 5.2. Two rather different communities

For each species in C_1 the probability of *not* being chosen on a given occasion is clearly

$$(1 - 1/11) = 0.901$$

The probability of not being chosen 11 times in a row is the compound probability,

$$0.901^{11} = 0.318$$

Since there are 11 species, the total contribution to their nonappearance, so to speak, must be

$$11 \times 0.318 = 3.494,$$

leaving a total expected number of species of 11 - 3.494 = 7.506.

One may apply the same analysis, albeit in a somewhat more complicated fashion, to the community C_2 as follows. The probability of species of abundance k not showing up in the sample is (1 - k/110). After 11 observations the probability has dropped to

$$(1 - k/110)^{11}$$

If one now merely adds up the contribution to non-appearances at each of the other abundances, the calculation takes only a few minutes.

$$\sum 2(1 - k/110)^{11} = 12.363$$

In this case the expected number of species in the sample must be 20 - 12.363 = 7.637. In other words, most of the time the two communities would yield the same number of species when sampled. Without knowledge of the distribution prevailing in C, the field biologist may make any prediction he or she likes but will be, under one scenario or the other, wrong.

It must be remarked that the shape of the samples of the two communities would differ somewhat, with C_1 having a higher initial peak than C_2 and declining somewhat more slowly, both petering out by the fifth or sixth abundance category. A more telling demonstration of the necessity of knowing the distribution of abundances in the community being sampled will be found in Section 9.2.1.

With a sampling intensity and appropriate distribution in hand, the next step in estimating the richness R of the community is to reconstruct C by estimating its parameters. Before presenting the two main techniques of richness estimation based on logistic-J theory, I will review some estimation schemes that have already been proposed.

5.2.1 Inadequacies in current methods of estimation

The natural interest in "biodiversity" in general and the need to know, specifically, the richness of communities, have led ecological theorists to seek methods of estimating richness in communities. The methods have proliferated, just as have the proposals for species distributions, as explained in Section 1.2, and the definitions of biodiversity in Section 3.3. The proliferation of methods takes place, as in the other areas, against a background that involves two central problems that presently face the methodology used in the past: a) a lack of adequate contact of theory with data and b) a failure to recognize that "non-parametric" approaches, those that fail to consider the distribution underlying the community as a whole, cannot succeed in any meaningful way, as illustrated in this section

The following review is hardly that. I have selected several of the more widely used methods to make the point just mentioned. Other studies could have been substituted for these without any change in the basic conclusions. The bottom line of this section is that none of the methods examined below could be called "exact" in the spirit of this book. Every method of which I am aware can only be called "approximate" and, indeed, prone to making false predictions.

Previous attempts to construct formulae or procedures for determining the number of species began with R. A. Fisher's derivation of the log-series distribution from the negative binomial. (Fisher et al. 1943). Subsequent developments (See (Chao 2005) for a review of methods.) were heavily influenced by this seminal paper, as we shall see.

The Fisher-Corbet-Williams method

In deriving his species estimation formulae, Fisher assumed that the parent community follows the normal ("Eulerian") distribution. This assumption stands in direct contradiction to the theorem that the species-abundance distribution of a sample must follow the distribution of the community it was drawn from. Field histograms never resemble the normal distribution and, thanks to the general theory of sampling, it can be said with some confidence that community distributions are never normal. Nor are they even approximately normal.

In any event, Fisher developed two equations that allow one to fit a log-series curve to a field histogram:

$$R' = - \alpha \ln(1-x), \text{ and}$$

$$N' = \alpha x/(1-x)$$

Here, as elsewhere in this book, R' and N' represent the number of species and individuals, respectively, in the sample. The parameters α and x are explained in Appendix A.5. Here we observe simply that when the sample values for R' and N' are substituted into these equations, corresponding values for the parameters a and x can be extracted by solving the equations. This was a troublesome matter in the 1930s when this work was done, but computers today solve these equations instantly. With values for α and x in hand, one has the theoretical form of the distribution thought by the authors to describe the sample.

Fisher went on to claim that, since the parameters α and x seem to remain constant over different sizes of sample from the same community, the relation between R' and N' can be extrapolated. Thus, with a higher value of N', a higher value of R' will result. This conclusion is largely conjectural, owing to the fact that only a few communities were examined in any detail. In any case, one may combine the two equations and write,

$$R' = - \alpha \ln(\alpha x/N') \qquad \text{(i)}$$

Here, as N' increases, the quantity $\alpha x/N'$ decreases. However, because the argument of the logarithm lies between 0 and 1, the logarithm itself is negative and increasingly so as N' increases. The minus sign makes the entire expression positive. Essentially, the expected number of species increases as the logarithm of the sample size. It is not clear how far the authors thought the method might be pushed in moving toward an estimate for R, the richness of the community as a whole. Since Fisher thought the community had a normal distribution, he would probably hesitate to suggest such a role for his formula. In his view of the matter, the log-series shape of sample histograms would necessarily morph into normality as N' approached N (the size of the community), invalidating the method. But given the closeness of the log-series distribution to the logistic-J distribution, at least at the low-abundance end of the axes, equation (i) might be used as a rough guide to the manner in which species accumulate as sample size grows. In the context of the Fisher, Williams and Corbett study, it would have been more consistent to assign the log-series distribution to a role in communities, as well as samples. However, any log-series distribution for the community would have much lower values for α than prevailed in samples, so an extension of the method would have to incorporate steadily decreasing values for α.

The Goodman statistic

Leo Goodman (1949) proposed the following estimator for the number of species in a

community having N individuals and a distribution F' of species over abundances in a sample. Goodman used an intermediate quantity S' in his formulation, here simplified to

$$S' = N - (N/n)F'(2),$$

where n is the sample size and F'(2) is the number of species having abundance 2. Goodman's estimator R may then be defined as follows:

$$R = S', \text{ provided that } S' \geq \sum^{n} F'(i)$$

$$= \sum^{n} F'(i) \text{ , provided that } S' < \sum^{n} F'(i)$$

It will be noted immediately that the summation can be replaced by R', the number of species to show up in the sample, so that one has the much simpler recipe,

$$R = S' \text{ if } S' \geq R'$$
$$= R' \text{ otherwise}$$

Goodman claims that although the statistic above is unbiased, it may only be applied when the sample size n equals or exceeds the largest population in the community. In Example One, described earlier, two markedly different communities C_1 and C_2 shared the same value of N. Moreover, when a sample of size 11 was selected from both communities, the same number of species tended to appear in the sample, namely 7.

Using the program SampSim (See Appendix A.), I sampled both communities 100 times at intensity r = 0.1 and obtained the following average values for F'(2) and S':

Community	F'(2)	S'
C1	2.11	88.9
C2	1.59	94.1

Table 5.2. Values of F'(2) for the two samples

Since S' > R' in both cases, the Goodman formula estimates 88.9 species in C1 and 94.1 Species in C2, both estimates being rather far off the actual richness values of 11 and 20, respectively.

In fairness to Goodman, it must be mentioned that his formula applies only to samples without replacement. On the other hand, one may suspect that the communities C_1 and C_2 may be scaled

up to the point where it makes little difference whether one samples with or without replacement. Any failures in this method would clearly be due to not taking into account the distribution that prevails in the community.

The Jackknife estimator

When a survey is taken of a population P of an animal species, a field biologist may capture an individual, mark it in some manner, then release it. The aim of this particular survey is to count the population. Allowing for the possibility that some individuals will be less prone to capture than others, one allows for varying probabilities of capture, say p1, p2, . . . , pm, where m is the population size.

A well-known method due to Burnham and Overton (1979) has been employed under these conditions to estimate the population P. Assuming that the individual capture probabilities are unknown, the method assigns a probability drawn at random from the interval [0, 1], the distribution being uniform.

The authors noticed that the problem as stated could be reinterpreted in the context of sampling a community for its species. The similarity can be nicely illustrated by the standard urn model of probability. The Urn A contains N balls marked with various numbers, some receiving the same number, some not. In fact the numbers are distributed among the balls according to the uniform random distribution just mentioned: if n(i) represents the number of balls receiving the numeric label i, then let F(k) represent the number of labels i that have k balls in their respective groups. The labeling can be arranged so that for each i,

$$n(i)/N = pi$$

Clearly, if one samples Urn A with replacement at each turn, the probability of drawing a ball from the ith numeric class equals the probability of capturing the ith individual in the population survey just described.

However, it may also be interpreted as the problem of sampling a community of species, with the ith species being represented by all the balls labeled i. If one were to make a histogram of this particular community, it would look basically level, with the usual statistical variations. The jackknife method is applied not only to the mark/recapture problem, but to the problem of sampling communities of species. However in the latter context, it clearly assumes a uniform distribution within the community and this cannot be correct, since it contradicts the theory of sampling (See Section 3. 6)

It may be that the jackknife method can be reformulated for the logistic-J distribution. If so, it would be hoped that the high variances that accompany the method do not exceed the variances of the richness estimation methods explored in Sections 5.3 below.

The Bootstrap method

Perhaps the main motivation behind non-parametric methods is to avoid having to know the distribution that prevails in the community, the underlying notion being that this information is present to some degree in the sample. The idea is clever, but subject to the "degree" of the "presence". As will be shown below, the bootstrap method has problems with certain distributions, a circumstance that undermines the idea of it being non-parametric.

The method, due to (Smith and van Belle 1984), uses a sample of n individuals to produce new samples by sampling the sample, so to speak. Since the sampling takes place (with replacement) at intensity 1.0, it amounts to a perturbation of the sample, as defined in Section 1.6.1. If the ith species shows up p_i times (relative frequency) in the perturbed sample, the following *bootstrap formula* is claimed to estimate community richness:

$$R = R' + \sum (1-p_i)^n,$$

Here, the summation is taken over the species present in the sample and R' is the richness of the sample. The process can be repeated, with a final estimate calculated as the average of estimates thus developed.

The term $(1-p_i)$ represents the probability that the ith species does not show up in a single drawing, whereas the term $(1-p_i)^n$ represents the probability that it does not show up in the sample at all. It will be noted that the same kind of probabilistic calculation was employed in the example that heads this section.

If p_i is small enough, the term $(1 - p_i)^n$ may not be negligible and the contributions from all such terms to the summation would be substantial. In other words, according to the formula above, the summation itself amounts to an estimate of the number of species that didn't appear in the sample; the low abundance species are used to estimate the number of missing species. For a sample size of 100, for example, a species of abundance 1, 2, or 3 in the sample will contribute the quantities 0.366, 0.133, or 0.048, respectively, the contributions from higher abundances decaying rapidly to zero.

By not taking intensity into account, the method fails, as the following counterexample shows.

Example Two

Let C_1 be the logistic-J community LJ[2.0, 1000] x 50 and let C_2 be another, smaller logistic-J community that happens to be an idealized sample of C_1 taken at intensity r = 0.1. When applied to C_1 the program Samplesim will produce such a sample by averaging over many samples, all

taken at the same intensity. This sample may also be considered as a logistic-J community in its own right, and turns out to have an average of 36.2 species. The same program can now be used to sample C_2 in the same manner, but this time at intensity r = 0.5. The final result is a sample that turns out to have an average of 29.5 species.

The point of this example is that the resulting (average) sample could have been obtained by sampling C_1 at intensity 0.05 or by sampling C_2 at intensity 0.5. The sample has 29.5 species, while C_1 has 50 and C_2 has 36.2 species. The bootstrap method cannot make both predictions simultaneously nor is it likely to make either, since there is a potential infinity of examples, all based on the same original distribution. Moreover, the same kind of example can be produced for any starting distribution one likes. I have used the logistic-J distribution for this example, since my software is geared to it. By using average samples I have avoided the possibility of producing an example on which the bootstrap method fails by coincidence. The averages used in the bootstrap method itself avail it nothing. An average of several incorrect answers is unlikely to be correct.

5.3 Exact estimation methods

In spite of the critical nature of the distribution underlying the abundances in a community, the following sampling formula plays no favorites, but applies to all possible (continuous) distributions that might lurk in the community. As explained in Section 3.6, the Pielou transformation represents an accurate template for the sampling process. It expresses the number F of species of abundance k to show up in a sample of intensity r:

$$F(k) = \int_0^\infty (e^{-rx}(rx)^k/k!).g(x)dx, \qquad \text{(ii)}$$

where g(x) is the distribution of abundances in the community being sampled. In the context of this book, the function g will always be logistic-J. (The reader may recall the claims made in section 3.6, earlier, that justified this assumption.) The Pielou transform is based on standard sampling theory as explained by Feller (1968). The formula is an extremely close approximation (error less than 0.1 %) to the hypergeometric formula on which the theorem is based. Indeed, the formula is also implicitly present (although not explicitly used) in one of the earliest papers on sampling in the ecological literature (Fisher, Corbet & Williams 1943).

It is not difficult to embed formula (ii) above in a computer program that takes an arbitrary (community) distribution as input and produces expected samples for every possible value of r, the sample intensity. (See description of the program CommRich in Appendix B.2) These outputs turn out to be essentially identical with the outputs of another program that simulates the sampling process itself in a statistically precise way. The program produces the expected number of species for each abundance category in a sample of given size (or value of r). This program is used to evaluate two methods for estimating the richness of a community from

samples of it. In doing so, it illustrates a general method of proceeding in estimation research.

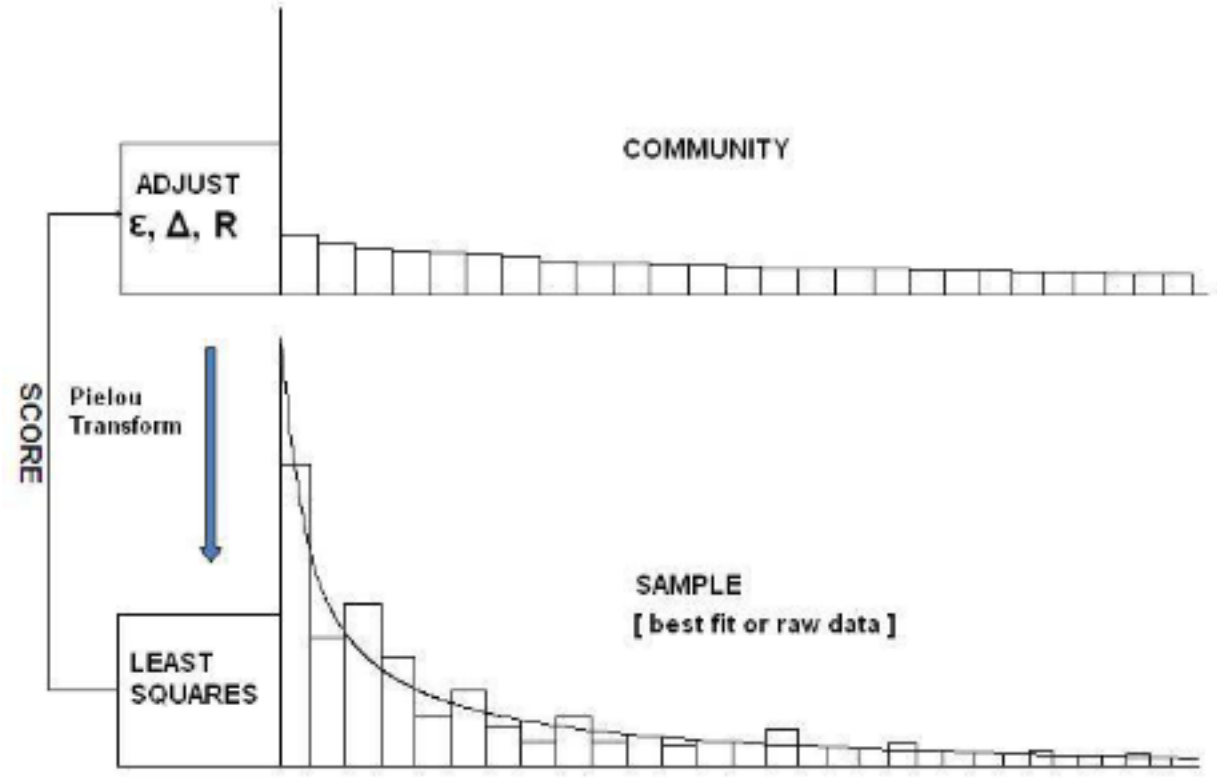

Figure 5.3. The estimation process converges to a best estimate for R

Both methods employ the same basic procedure of cycling back and forth between sample data and a test-community LJ[e, D] x R, as shown in Figure 5.3. The community is sampled iteratively, always at the same given intensity r, but each time with different values for ε, Δ, and R, until a certain combination of the parameter values produces a sample that most closely matches the sample in hand. The iterative process uses the method of steepest descent, always moving in the direction of an improved match and arriving there after a finite number of steps.

The sample drawn from the community at each stage of the process is an *expected sample* arrived at via the Pielou transform, as in Figure 5.3. In the expected sample each abundance category is inhabited by a fractional number of species, such as 12.64 or 0.42, etc. These numbers have the logistic-J distribution and they may be compared directly with their counterparts in the sample at hand. In the two-step method the sample is computed as the best fit to the field histogram. In the one-step method, the sample is the histogram itself. The least squares function provides the vehicle for the comparison, where F”(k) is the expected value for the number of species in the kth abundance category and F’(k) is the actual number of species in the sample at hand (whether raw or computed) which have that abundance.

$$\text{difference} = \sum(F''(k) - F'(k))^2,$$

The basic cycle may be described as follows:

1. Give initial values to the parameters ε and Δ for the community, as well as the community richness, R.

2. Use the Pielou transform to produce the expected sample of this community.

3. Compare the resulting theoretical sample with the one in hand via the least squares measure.

4. If the match is worse than before, reset the most recently changed value for ε, Δ, or R in the opposite direction. If the match is better, continue as before.

Presently, the entire process is embedded in one of two computer programs, depending on the method, and the complete richness estimation process may take anywhere from 2 to 200 cycles to complete. In other words, a human must execute the search algorithm, a process that can itself be automated, cutting the estimation time from an hour or two down to a millisecond. The algorithm itself systematically cycles through ε, Δ, and R, changing each until no further improvement is seen, then switching to the next parameter. At no point do the corresponding parameter values for the sample at hand play any role. The sample richness R' plays an implicit role, however, through the values of the sample function F' at each abundance category.

5.3.1 The two-step method with an example

The first procedure described here is called the *two-step method*. It proceeds in two main steps: Step 1. Find a best-fit for the sample histogram with the logistic-J distribution. The program called BestFit does this, taking the sample histogram as input, then comparing these data with the numbers generated from a theoretical (logistic-J) sample distribution with (sample) parameter values input by the user of the program. The values of ε' and Δ' thus arrived at can be varied systematically over the parameter space to discover a global minimum in solution space. In most cases the method of steepest descent finds the minimum without having to search the entire space. The measure of fit is the chi square score divided by the number of degrees of freedom, as determined by the program. This method of scoring helps to minimize jumps in score values that would otherwise result when the program changes the number of degrees of freedom.

Step 2. One then inputs the best fit logistic-J parameter values into the program CommRich, along with the sample intensity estimate, r, made by the biologist. The user then conducts a directed search through solution space by systematically varying the community parameter values ε and Δ, as well as the community richness R, as described above; for each set of values thus arrived at, the program computes values for the expected sample and compares the

theoretical sample with the best fit curve from Step 1 using the least squares formula as a measure of similarity. The underlying algorithm uses the smallest least squares score found so far as the basis for further improvements in the score. Any change in a parameter value that leads to an improved score is adopted as the starting point for the next step. The change is not selected arbitrarily, but on the basis of producing the greatest improvement of the score, as it steadily descends toward zero. At the end of the convergence process one reads off not only statistically accurate estimates for ε and Δ in the community, but its richness, R, as a byproduct of the process.

Time to convergence during either fitting process depends strongly on the starting parameter values. But a form of binary search may be employed that speeds the process up, completing in a time that is proportional to the logarithm of the size of the parameter space being searched.

An example of the method in action is provided by data sent to me by M. G. M. Jansen, a Dutch biologist who has been conducting an extensive sampling program for lepidoptera inhabiting coastal salt marshes in the Netherlands. Table 5.3 displays the data from one of Jansen's samples. Each cell of the table under the heading "no. spp." also contains the corresponding number of species predicted for the corresponding abundance. The table shows observed abundances for some 45 species, the remainder having abundances 31, 32, 41, 67, 103, 103, 1121, and 2073.

abund.	**no. spp.**	**abund.**	**no. spp.**	**abund.**	**no. spp.**
1	15 15.24	11	1 0.87	21	1 0.42
2	7 5.75	12	0 0.79	22	0 0.40
3	3 3.60	13	1 0.73	23	0 0.38
4	3 2.62	14	0 0.67	24	0 0.36
5	2 2.05	15	0 0.62	25	1 0.34
6	1 1.68	16	2 0.58	26	0 0.33
7	2 1.43	17	0 0.54	27	0 0.31
8	2 1.23	18	0 0.50	28	0 0.30
9	1 1.09	19	0 0.47	29	0 0.29
10	1 0.97	20	1 0.45	30	1 0.28

Table 5.3. Sample abundances vs predicted ones

Running the program *BestFit* on these data, I found a best fit for the chi square measure at parameter values ε' = 0.358 and Δ' = 124.7. The resulting chi square score was 0.5372 at 6 degrees of freedom. The optimum fit was obtained by starting at initial values of ε' = 0.5 and Δ' = 140.0 and continuing to adjust the values of the two parameters, steadily decreasing the chi square score, until no further improvement was possible, as described in Section 4.3 in the previous chapter. The procedure required 22 steps in this case.

With these values of ε' and Δ' as input, the program CommRich produced a theoretical version of the sample, storing it in an array for the ensuing cycle of comparisons. The first 30 parallel values arise from the logistic-J distribution LJ[0.358, 124.7] x 54.

Jansen's estimate of sample intensity, namely r = 0.0017 (0.17 percent of the community sampled), was inputted to the program CommRich. I then followed the method of steepest descent, starting at initial guesses of e' = 1.0, D' = 2200.0 and R = 100. It took some 60 steps to arrive at values for all three parameters that collectively minimized the least squares measure. The values were e' = 2.50, D' = 2215.0, and R = 127.

Jansen thought the R-estimate rather high. He would have estimated the number to be closer to 100, based on intuition about salt marsh lepidopteran abundance in general. However, he had shown some uncertainty about his estimate for the intensity, r, of his sample. The reasons for his uncertainty lay in the nature of his subject. The salt marsh under study had some variety in its vegetative structure and some species seem to be more abundant than samples would indicate, as though sample intensities were themselves varying. In any case, for a higher value of r, the estimate for R would have been lower. On the other hand, biologists are often surprised at communities that turn out to be more speciose than they thought.

How good is the R-estimate? The answer is provided by a series of experiments described in the next section. In the case at hand the estimated number of species in Jansen's lepidopteran community was 127 species, give or take 6 percent (8 species) 95 percent of the time. In other words, with probability 0.95 the lepidopteran community (would have) had between 119 and 135 species -- at least if the r-value supplied by Jansen was approximately correct. If his estimate of the sampling intensity r were low, however, and if the actual value were 0.0020, for example, the R-estimate would drop to 106 species.

All of this raises the obvious question of how we can ever know the number of species in a community that is under study. The simple answer is that unless we sample everything, we can't. However, the exact methods described in this book will produce estimates with a statistical accuracy that, in the author's opinion, cannot be improved. Two kinds of communities are used as test beds for the estimation procedures. One kind is purely theoretical, being a strict logistic-J distribution with the parameter values currently under test. The other kind is a perturbation of such a community, resulting in a distribution that I will claim to be typical of the distributions that prevail in real communities.

5.4 Experimental illustration of methods

The following experiments are not used to prove that richness estimation methods work. In the context of LJ theory they are already known to work. The experiments below serve merely to a) illustrate the accuracy of the method and to b) provide interval statistics (based on variance of the samples) so that reliable error estimates of accuracy can be made. This means that each estimate made by this method will have an accompanying error term.

A second set of experiments was piggybacked onto the first set. By using two versions of the community being sampled, it was possible to arrive at preliminary estimates of the relative importance of variation in community abundances versus variation in those of the sample. Two forms of the community being sampled were used. In one of these the community was represented by a smooth theoretical function. In terms of logistic-J theory, such a community will have zero variation. The second form of community was a perturbed version of the first.

A total of 75 tests were carried out, each involving a random sample drawn from an LJ community using the computer program called SampleSim.PAS. The experimental design is summarized in Table 5.4

method:	**two-step**	**one-step**
idealized:		
perturbed:		

Table 5.4. Basic 2 x 2 experimental plan

5.4.1 The two-step method

In the first experiment 25 random samples were drawn from an idealized community LJ[2.0, 3000.0] x 50. In the second experiment 50 random samples were drawn from randomly perturbed versions of LJ[2.0, 3000.0] x 50. These were obtained by simply sampling (with replacement) the community at intensity 1.0. The results of these experiments are summarized in Tables 5.5 and 5.6. In each case the average estimated richness appears under the heading “mean”.

r-value	**# tests**	**sample size**	**mean**	**s.d.**	**error**
0.005	10	68	50.47	7.02	8.3%
0.010	15	136	49.50	3.47	5.0%

Table 5.5. Richness estimates and error terms for idealized community

r-value	# tests	sample size	mean	s.d.	error
0.005	12	68	46.9	10.4	13.6%
0.010	15	136	50.3	10.6	12.5%
0.020	14	271	49.4	5.7	6.8%
0.020	10	387	51.4	3.9	5.6%

Table 5.6. Richness estimates and error terms for perturbed community

The two average richness predictions that appear in Table 5.5 are closer to the ideal average of 50.0 than are the average predictions shown in Table 5.6. This is due to the data being less variable, given that samples came from a smooth, idealized community. At the same time, we observe that in both tables the error term for intensity 0.005 is larger than the error term for intensity 0.010. This makes sense because the samples are twice as large in the second case.

The results displayed in Table 5.6 show a similar trend in accuracy; the error estimate declines in step with higher sampling ratios. The last set of (10) samples in Table 5.4 were drawn from a much "larger" community, a perturbed version of LJ[10.0, 4000.0] x 50. This not only illustrates the method's wide applicability, but provides an extra data point when plotting estimation error as a function of sample size. (See the end of Section 5.) The average richness estimate for all four tests is 49.5, acceptably close to 50.

5.4.2 The one-step method

In this method the second step of the two-step method is used in the same manner, but with a different input -- the raw data itself, instead of the best fit. There is, therefore, no first step to speak of. Tests of the one-step method parallel those of the two-step method, with the first series of experiments performed on the idealized community LJ[2.0, 3000] x 50. The remaining tests focus on perturbed versions of the same community. The series two tests thus assess the accuracy of the one-step method, enabling a direct comparison of the results.

r-value	# tests	sample size	mean	s.d.	error
0.005	10	68	53.0	8.9	10.4%
0.010	15	136	49.1	4.0	5.7%

Table 5.7. Richness and error estimates for idealized community

r-value	# tests	sample size	mean	s.d.	error
0.005	14	68	52.0	10.9	13.2%
0.010	14	136	51.9	10.3	12.6%
0.020	14	271	51.4	7.0	8.4%
0.020	10	387	51.4	3.4	5.7%

Table 5.8. Richness and error estimates for the perturbed communities

Results of the first series of experiments on the idealized community LJ[2.0, 3000] x 50 are summarized in Table 5.7, while those for perturbed versions appear in Table 5.8.

Apart from what appear to be significantly higher error terms in the one-step method, there is little to choose between the two methods. Given that the mean richness estimate is independent of intensity, one can arrive at a slightly more refined richness estimate of 49.98 % for the two-step method and 51.05 % for the one-step method, both applying to the idealized community.

The foregoing results arise from the idealized community, of course. The second set of experiments involved not an idealized community, but a great many perturbations of it. In this case, the richness estimates averaged 49.5% for the two-step method and 51.7% for the one-step method. The second average seems a bit high, but the 50% target is well within the 8.4% error interval, [47.4, 56.0].

5.5 The behaviour of error terms

Although more random communities could be sampled in order to obtain more precise data for plotting error as a function of sample size, the results of the previous section yield a reasonably close first approximation. Table 5.9 illustrates how expected errors could be systematically derived from the experiments described in this chapter. The table is based on the error curve derived from the experimental data.

n	60	80	100	120	140	160	180	200	220	240	260	280
error	15.9	13.9	12.4	11.4	10.5	9.9	9.4	8.9	8.5	8.1	7.9	7.6
n	300	320	340	360	380	400	420	440	460	480	500	520
error	7.3	7.1	6.9	6.8	6.6	6.4	6.3	6.1	6.0	5.9	5.8	5.8

Table 5.9. Practical table for the assessment of errors in richness estimation

The error curve shown in Figure 5.4 is an optimum fit to the error data shown in Tables 5.5 to 5.8, using a function of the form $y = k/\sqrt{x}$, where x is the sample size. This is a statistically standard application, error declining as the inverse square root of sample size.

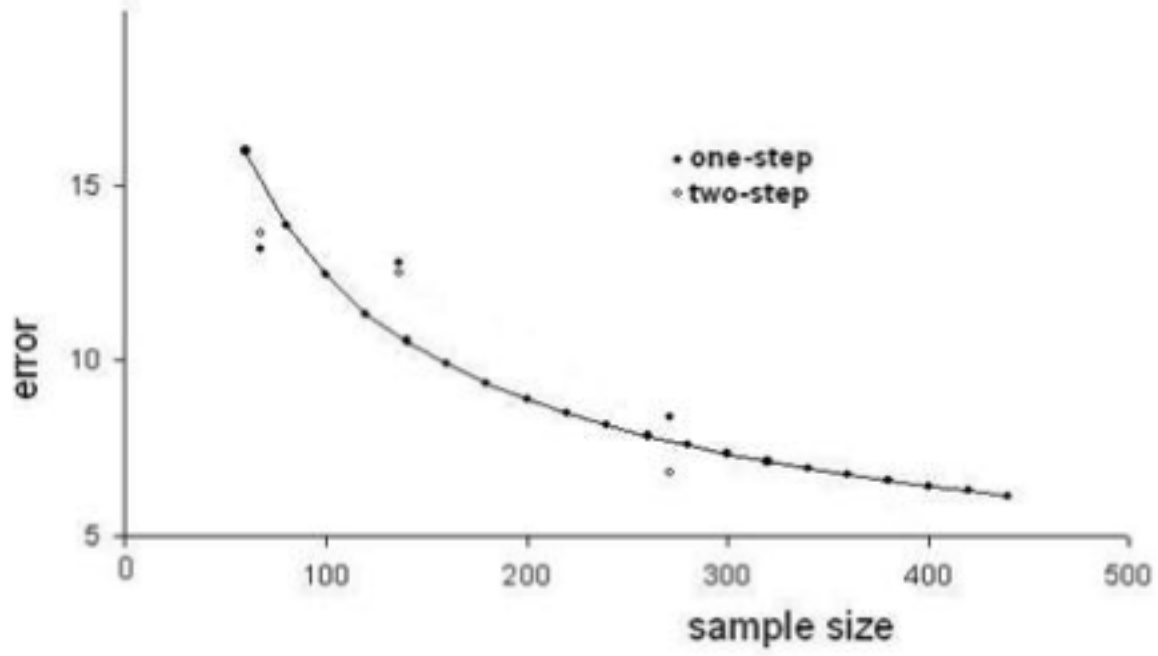

Figure 5.4. richness estimation error as a function of sample size

As will be seen in the next section, the percentage contribution to variance by the source community is relatively low for small sample sizes, with (pure) sample variance predominating. The resulting error curve is therefore the weighted sum of two functions, the weights gradually shifting from predominantly sample-influenced to community-influenced, as sample size increases. The approximation curve is reasonably close to the error data and, particularly for larger samples, may overestimate the error.

As can be seen from the relative positions of open (one-step) and closed (two-step) circles, there is little to choose between the two methods in respect of accuracy. The entries in table 5.9 were read directly from the curve, in effect. The table, although useful in a limited way, serves only as an example until further experiments sharpen and extend these results.

5.6 Analysis of the two sources of variance

As was pointed out earlier in this chapter, both the community and the sample contribute some variance to the final estimate of richness. The point of the Series One tests was to tease out the variance due to the sample alone. The results of Series Two incorporate both sources of variation and they are, of course larger. Corresponding entries of Tables 5.7 and 5.8 may be compared to

analyse the respective contributions at two intensities or, in this case, sample sizes.

Idealized communities

ratio r	samp. size	# tests	two-step: variance	one-step: variance
0.005	68	10	49	79
0.010	136	15	12	16

Perturbed communities

ratio r	samp. size	# tests	two-step: variance	one-step: variance
0.005	68	10	108	118
0.010	136	15	94	106

Table 5.10. Contributions to variances in methods from smooth vs ragged communities

Since the two sources of variation are completely independent, the variance in the sample may be calculated as a percentage of overall variance:

For the two-step method the proportion (percentage) of independent sample variance is about 45% at sample size 68, dropping to roughly 13% at sample size 136. For the one-step method, the proportion falls from approximately 68% to 15%. It makes sense that the proportion should drop since larger samples tend to have less relative variation than smaller ones for the same source community, whatever its state. On the other hand, the one step method produces larger percentages across the board, the difference diminishing as the sample size increases.

There is a clear tradeoff between the two methods. The one step method is easier to carry out and enjoys much the same accuracy as the two-step method, but it produces a more diffuse kind of prediction, for which the error term is significantly larger than that of the two-step method.