## Chapter 4. Compiling and analysing field data

No experienced field biologist needs to be told what to do in compiling a list of species and abundances for a specific area. Yet the same biologist may be unaware of what biologists in quite different areas of biodiversity assessment may be doing. In fact, there is a general pattern or template that can be applied to all sampling activity of this kind. Laying out the template gives me the opportunity of linking the general activity with the specific requirements of the logistic-J distribution.

When sampling in the field one invariably keeps records, whether the organisms being sampled are identified on the spot, through photographic records, DNA samples, or taken back to the lab. Normally the number of individuals within each species are counted. The raw counts may be used directly or converted to densities or percentages on an areal or volumetric basis. The logistic-J distribution is equally friendly to all forms of count or count surrogates.

In what follows I describe basic methodology for the arrangement and display of field data, as well as the two most widely used measures of fit for theoretical distributions versus field histograms. A method for estimating more accurately the true species overlap of two samples appears at the end of this chapter as an illustration of the potentially wider role to be played by the logistic-J distribution in the analysis of data.

### 4.1 Histograms and Distributions

A histogram is a compilation of data into categories for the purpose of revealing a shape or trend that might not be obvious from examining the data as a mere list of numbers. In the case of abundances appearing in a sample, the most natural categories are ranges of values into which the abundances may be sorted. Here, for example, is a set of abundances of Microlepidoptera taken in a light trap in The Netherlands in 2006 (Jansen 2008). The order of abundances is determined by the order of the respective species, taken in some canonical order, as in taxonomic synopses, or alphabetical order, or no particular order, the point being that the abundance counts (number of individuals per species) are usually *not* in order:

> 1121, 6, 2, 2, 2, 10, 2, 8, 32, 31, 16, 21, 67, 9, 5, 103, 25, 1, 2, 1, 1, 1, 1, 13, 1, 2073,
> 1, 20, 41, 1, 3, 5, 3, 7, 103, 2, 1, 136, 1, 1, 1, 1, 4, 2, 1, 11, 3, 7, 1, 4, 4, 30, 16, 8, 1

Before constructing a histogram one must decide on categories, typically embracing a range of values. For present purposes, we will adopt the simplest, unitary categories such as 1, 2, 3, etc.

There are several methods for compiling a histogram from data like these. The most direct is to count the species having abundances that fall into each category, as in Table 4.1 below:

| Category | # spp. | Category | # spp. | Category | # spp. |
|---|---|---|---|---|---|
| 1 | 16 | 11 | 1 | 21 | 1 |
| 2 | 7 | 12 | 0 | 22 | 0 |
| 3 | 3 | 13 | 1 | 23 | 0 |
| 4 | 3 | 14 | 0 | 24 | 0 |
| 5 | 2 | 15 | 0 | 25 | 1 |
| 6 | 1 | 16 | 1 | 26 | 0 |
| 7 | 2 | 17 | 0 | 27 | 0 |
| 8 | 2 | 18 | 0 | 28 | 0 |
| 9 | 1 | 19 | 0 | 29 | 0 |
| 10 | 1 | 20 | 1 | 30 | 1 |

**Table 4.1.** Tabulation of counts by number of species per abundance

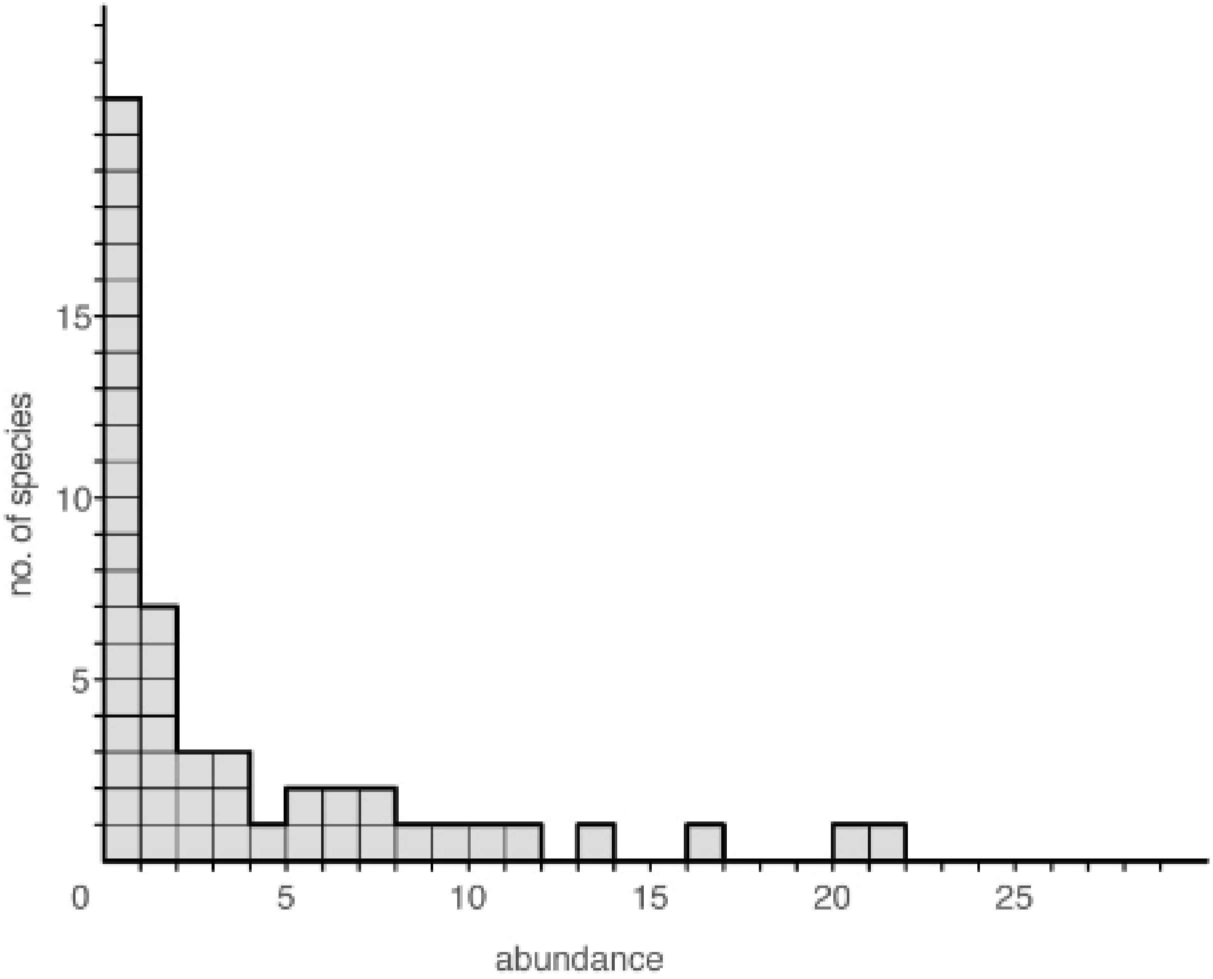
The invariable result of such a compilation is that more and more categories with a zero entry appear as the abundance variable increases. At some point, best determined by the analyst, it becomes simpler and more economical to list the remaining abundances in increasing order. In the case at hand, one would then simply list the abundances not tabulated.

31, 32, 41, 67, 103, 103, 136, 1121, 2073

The foregoing table can be converted into a graphical histogram (bar chart form) by creating a column for each category. The height of the column will be proportional to the abundance entry for that category. Figure 4.1 shows such a histogram of this type derived from the data just compiled numerically.

The two species sharing an abundance of 103 may strike some readers as improbable, as indeed it is. However such statistical coincidences are almost certain to occur in any reasonably large dataset. The probability of two species having the same abundance is close to 1 at abundance 1 and falls off gradually for ever-higher abundances.

This histogram in Figure 4.1 reveals the typical J-shape of field data, albeit with an unusually high unit category bar. In fact, although selected fortuitously and with no foreknowledge of how Jensen's data might appear in histogram form, the result is relatively smooth. Often such histo-

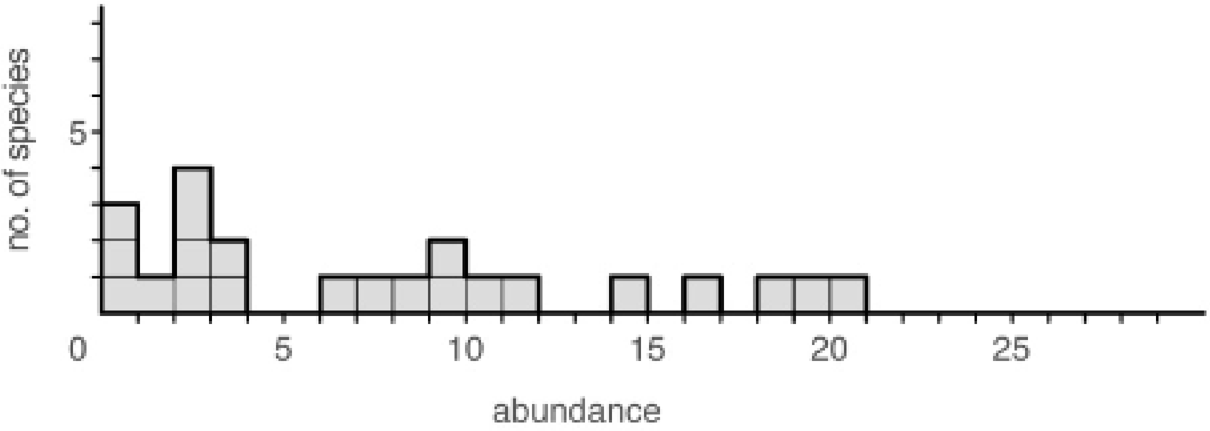


| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| incl. spp at: | 30 | 31 | 32 | 41 | 67 | 103 | 103 | 136 | 1121 | 2073 |

**Figure 4.1.** Example of a field histogram for the data in Table 4.1

grams are more ragged than this, especially as one moves to higher abundances. There, one enncounters species gaps of wildly varying widths, accompanied by two (or more) species piled up at one abundance. Indeed, the Jansen data happens to have two species of abundance 103.

Much field data plots as a relatively shallow J-curve, with species gaps occurring even at low abundances. In such cases it may be desirable to group the data by twos or threes. For example, the histogram of abundances shown in Figure 4,.2 represents numbers observed by Busby and Parmelee (1927) in their survey of herpetofauna in Kansas. It has the typically ragged shape of data in which fewer species, on average, inhabit the lower abundance categories.



| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| incl. spp at: | 25 | 28 | 29 | 35 | 37 | 39 | 52 | 57 | 105 | 142 | 197 | 206 | 353 | 437 | 617 | 713 |

**Figure 4.2.** Field histogram of Herpetofauna

When the categories are changed by grouping, the shape implicit in the data is more clearly seen. Here we have grouped the abundances by fives.

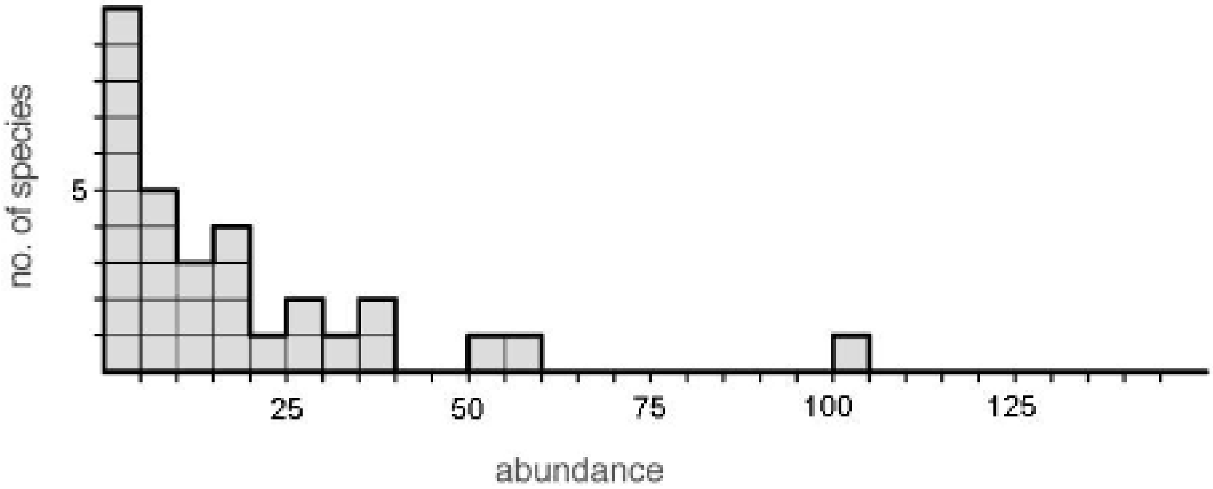


**Figure 4.3.** Histogram of Figure 3.2 after grouping operation [change numbers]

The remaining raggedness is normal, not only in unitary abundance categories but in grouped categories, as seen here. Often a field biologist will report abundance data in terms of densities of individuals per unit area. Densities are no different from whole number counts in being subject to compilation into categories. The logistic-J distribution, being continuous, works just as well with these data as with whole number counts. In such a case, the abundance axis might well be labeled with decimal numbers such as 0.1, 0.2, 0.3, etc.

## 4.2 Other Representations

In the previous sections of this chapter, indeed throughout this book, I have used what I call the standard axis suystem, with abundances playing the role of independent variable and numbers of species (richness within abundance categories) playing the role of dependent variable. This seems the natural place to point out that the standard axes are ideally suited to the logistic-J distribution, particularly in its dynamic aspects. Only in this axis system are population fluctuations so easily visualized as "vibrations" on the abundance axis. It is also ideally suited to the presentation of taxonomic abundance, as presented in Chapter Eight. There, we may substitute the abundances of genera, for example, for abundances of families, plotting the number of families per generic abundance, instead of the number of species per individual abundance. The spectrum of all such possibilities forms a seamless unity, with the standard axes appearing throughout.

### 4.2.1 Rank abundance

The *rank abundance diagram* is obtained by placing observed abundances in monotonic decreasing order, taking their logarithms, then plotting the results in bar form. For example, if one uses the abundance data from the first example of the foregoing section, take their logarithms (to the base e), place them in rank order and plot them, we obtain Figure 4.4.

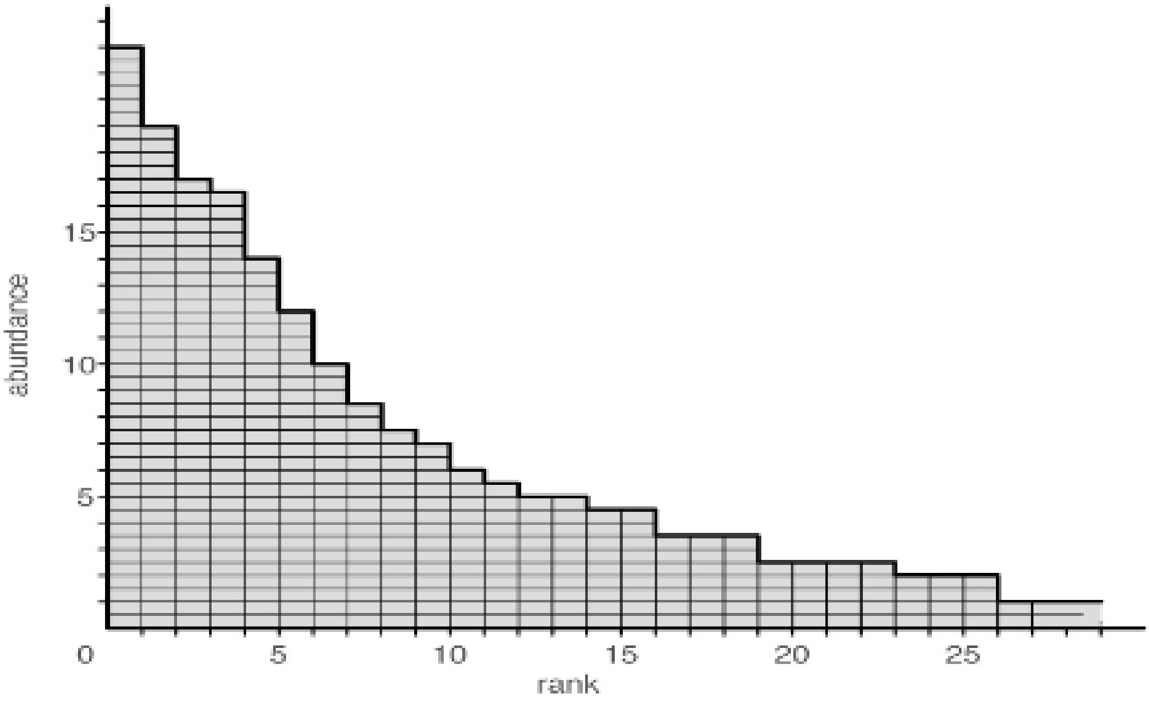


**Figure 4.4.** Rank abundance diagram for the data of Figure 3.2

Apart from the disadvantage of vertical scale compression, the rank abundance diagram is somewhat unwieldy mathematically speaking. The horizontal axis is not a metric axis, but an ordinal one, so metric operations such as grouping or scaling do not apply in any meaningful sense. Moreover, operations such as the insertion or deletion of a species, trivial to carry out in the standard abundance diagram, involve shifting all the species on the right hand side of the one deleted or inserted. In terms of the information stored in it, however, the rank abundance diagram is equivalent to the standard system -- which can be reconstructed by reversing the process.

It might be asked at this pojnt if the logistic-J distribution can be cast in a similar form. The answer turns out to be "almost". It happens that the logistic-J distribution in invertable, that is one ca exchange axes and the distribution has the same form. To invert the logistic-J pdf, one smply interchanges the variables x and y, along with their displacements:

$$x = c(1/(y+d) - e$$

The distribution function differs from the pdf in having the factor R included in the following manner: Instead or replacing x by y, replace it by y/R. The replacement of y proceeds as before In this inverted distribution, numbers of species label the horizontal axis and abundances label the vertical one, just as in rank abundance.

$$F(y) = c(1/(y/R + d) - e$$

The distribution just described, however, ranks the abundances in a different way. The best interpretation I can give to the fingure requires the following definition: a k-tuplet is a set of k species that all have abundance k where k is the maximum such abundance over all k-tuplets. Under this interpretation, F(k) represents the average abundance of a k-tuplet. This notion specializes to the value k = 1 which happens to be the average maximum abundance, D or the appropriate modification of it. It strikes on e that for the time being, this distribution is hardly more than a curiousity

### 4.2.2 Logarithmic Abundance

The *logarithmic abundance diagram* does not contain the equivalent information to the standard axis system. In this scheme, abundance categories are grouped into "octaves," the first octave consisting of the first or lowest abundance, the next octave consisting of the next two abundances, the octave following that one consisting of the next four abundances, and so on, with the kth octave consisting of the abundances $2^{k-1}$ to $2^{k}$-1. Plotted in this manner, the Jensen data appears in Figure 4.5. Normally such data has the appearance of a truncated hump without an initial spike, as here. I have labeled each category with the highest abundance in its group. If I used a base higher than 2, the hump may well appear.

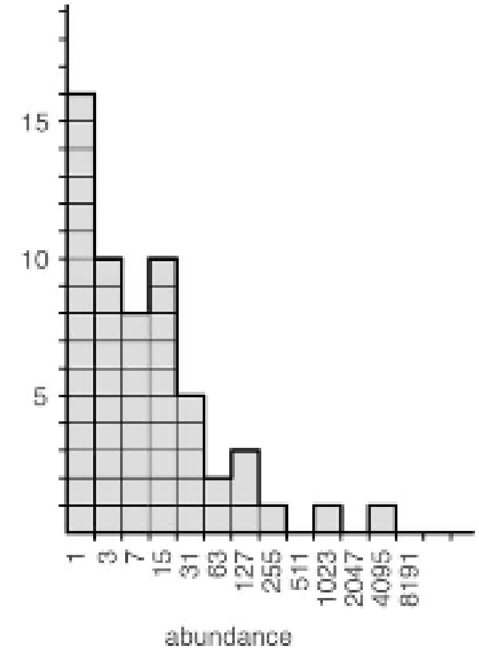


**Figure 4.5.** Histogram of data in Figure 4.2 grouped by "octaves"

Although the logarithmic abundance system has undoubted advantages in the inclusion of very large numbers, it pays a heavy price in information loss. This system is *not* equivalent to either the standard system or the rank abundance diagram. Information is normally measured in binary digits or bits, but decimal digits may be used to carry the point in the present instance. If one simply counts up the digits in the biodiversity vector of the first example, one obtains a total of 78 digits. But if one counts the total number of digits required to specify the logarithmic form of these data, one gets 15 digits. In this case, logarithmic grouping has destroyed about 81 percent of the information in the data.

The supposed charm of the lognormal distribution arises from the bell-shaped curve, usually truncated on the left, that emerges when species abundance data is plotted by octaves. Sometimes (as above) the shape does not emerge with any great clarity. As I explained at the end of the previous chapter, however, if one subjects the log-series distribution to a logarithmic transformation, a truncated unimodal curve emerges. (Dewdney 1998). Moreover the bell-shaped curves that appear under these circumstances are completely indistinguishable, at least by visual inspection, from those that arise from the lognormal distribution. In other words, the appearance of a bell-shaped curve in this context does *not* permit one to conclude that the data subjected to the transformation have the lognormal distribution.

Apart from the slight advantage of being able to represent higher abundances using either the log-transformed representation or the rank abundance diagram, there is no particular advantage in using either distribution. All field data and theoretical curves are well represented by the standard axis system. Indeed, it has the additional advantage of being the conceptual theatre in which we may visualize stochastic vibrations directly.

### 4.3. Estimating parameters of the logistic-J distribution

When the field biologist has taken one or more samples of the community under study, the first step in analysing the data is to plot the species abundance histogram as described in the previous section. The next step, although not necessary for all purposes, is to find the best fit of the logistic-J distribution to the data as plotted.

To find the best fit of field data to the logistic-J distribution, the easiest measure to use is the chi square goodness-of-fit test. For an exact best fit, there is no alternative to testing more than one combination of values for the parameters $\varepsilon$ and $\Delta$. A method for finding an optimum fit to the logistic-J is outlined in the next section. A reasonably good (suboptimal) fit may be obtained indirectly from the mean abundance $\mu$ and the height $F_a$ of the minimum abundance peak in the field data. The height is the number of species in the minimum abundance category a, whether an integer or a fractional number, as explained in the previous chapter. Transfer equations (Appendix A.2), taking $\mu$ and $F_a$ as input, may then be solved to yield $\varepsilon$ and $\Delta$ as output. One may then plot the fit developed by either method and compare it directly to the field histogram.

The transfer equations may be solved by a variation of Newton's method in which successive estimates for the two parameters converge to an exact solution. Although it is always possible to solve the equations by hand (with a calculator), a computer program is preferred. (See "Solveit" in Appendix A.10). Biologists who find this prospect daunting for whatever reason are invited to send their data to the author who can find optimal fits rather quickly. Later, by visiting the author's website, they will find online software that does this.

#### 4.3.1 The chi square test

When abundance data are compiled into a standard histogram it becomes possible to compare the resulting shape with various theoretical proposals, including the logistic-J distribution. However, comparisons based on visual similarities can be misleading and somewhat dangerous. Not just some but all of the abundance models developed prior to 1999 were "established" by this method, which may explain why there are so many.

An objective method of comparison is entirely quantitative and does not depend on subjective judgments. Carl Pearson, who developed the chi square goodness-of-fit test, used a special statistic to measure the difference between an actual abundance $a_i$ and a corresponding theoretical abundance $t_i$ :

$$d(a_i, t_i) = (a_i - t_i)^2/t_i$$

The sum of such differences is called the *chi square test statistic* and is compared with numbers in a chi square table to determine to what degree the empirical data matches the theoretical prediction. Each component in such a sum is considered as one “degree of freedom,” meaning that it contributes freely and independently to the overall sum. However, each parameter of the theoretical distribution under test amounts to a restriction on how freely the terms may vary. Pearson therefore subtracted the number of such parameters from the number of terms in the sum, yielding the *degrees of freedom* in the test being carried out. The logistic-J distribution has two parameters, ε and Δ, so that a test involving 12 abundance categories, for example, would have ten degrees of freedom.

In computing a chi square test, there is a requirement that when theoretical frequencies fall below 5.0, the corresponding consecutive abundance categories must be grouped so that their sum is ≥ 5. For example, if the expected abundances at 31 and 32 are 4.22 and 3.90, respectively, while the empirical abundances are 3 and 4, respectively, the corresponding term in the chi square statistic would be:

$$(7 - 8.12)^2 / 8.12,$$

where 7 = 3 + 4 and 8.12 = 4.22 + 3.90. Further out in the abundance axis, the last empirical abundance category with a nonzero entry may well exceed Δ. Whether or not this happens, the last (grouped) category will use the sum of all remaining theoretical values to be compared with the sum of all remaining empirical values.

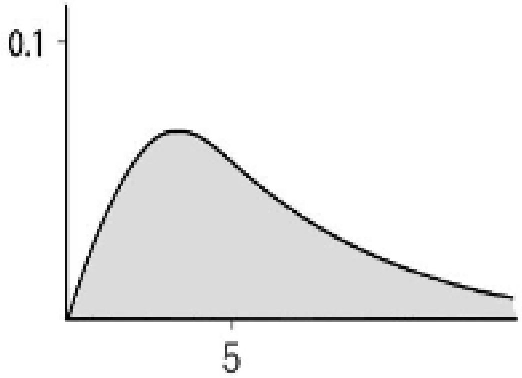


**Figure 4.6.** The chi square distribution with five degrees of freedom

Under the null hypothesis, that the data observed actually arise from or follow the theoretical distribution in question, Pearson showed that chi square test scores must themselves have a certain distribution, which he called the chi square distribution, as shown in Figure 4.6. In other words, if one carried out 100 chi square goodness-of-fit tests (with five degrees of freedom) on data which one knew beforehand originated with the same logistic-J distribution, one would find that the test scores, when compiled into a histogram, would effectively match this outline. As one can see from Figure 4.6, there will always be “outliers,” instances of the distribution that fit rather poorly and so give rise to a high score, possibly off the page in Figure 4.6. Such samples may even match another version of the theoretical distribution under test much better, yet they originated in the distribution at hand.

As normally applied, the chi square test invokes a null hypothesis that the empirical data arise from the theoretical curve being matched to it. The resulting score is compared to a table of critical values, numbers that mark the boundaries between acceptance and rejection of the null hypothesis at various levels of significance.

| **q** | 0.995 | 0.990 | 0.975 | 0.950 | 0.900 | 0.750 | 0.500 | 0.250 | 0.100 | 0.050 |
|---|---|---|---|---|---|---|---|---|---|---|
| **8 df** | 1.344 | 1.647 | 2.180 | 2.733 | 3.489 | 5.071 | 7.344 | 10.22 | 13.36 | 15.51 |

**Table 4.2.** Critical values for the chi square distribution

Suppose that a chi square test is carried out at 8 degrees of freedom on two different sets of empirical data, with resulting scores of 6.732 and 2.451 when compared to a particular theoretical distribution. Table 4.2 contains the critical values for eight degrees of freedom (Pearson & Hartley 1972).

The upper row contains a set of q-values. If a score exceeds a particular critical value, then the fit is rejected, with the corresponding probability q of being right. Thus the score 6.732, which is greater than 5.071, involves a rejection of the null hypothesis with a probability Q of 0.750 of being right. On the other hand, the null hypothesis would be "accepted" at the critical value of 7.344, but here the probability of a correct rejection drops to 0.500. In this instance "acceptance" simply means non-rejection. It does not mean that the data under test actually originated with the theoretical distribution at hand.

In normal use, the test statistic is compared with the critical value at the Q = 0.950 (95%) level of rejection;  one wants a fairly high probability of being right in rejecting an hypothesis. At this level, the score of 6.732 is greater than the critical value of 2.733 that corresponds to the 95% level and the null hypothesis is rejected, with a probability of 0.95 of being correct. However, the other test score of 2.451, being less than 2.733, is "accepted."

An important feature of the chi square test (and, indeed, most goodness-of-fit tests) is that it applies to any theoretical distribution whatever. It applies, of course, to the logistic-J distribution as well.

In Section 7.3.1 I shall be more specific in describing the relationship between the theoretical distribution and the histograms being matched to it. Under the null hypothesis one can be more specific about this relationship than saying that the data "arise from" or "follow" the distribution. In Section 7.3.1, I will describe the histograms as *perturbations* of the theoretical distribution (called the *source*), obtained by varying the theoretical column heights in a specific random manner, the result closely resembling a sample of the theoretical distribution at intensity r = 1.0.

### 4.3.2 Finding an optimum chi square fit

Sometimes a chi square test is applied simply, as in the case when the parameters of a theoretical distribution are known (or estimated) beforehand. But sometimes one must apply the test many times (to the same field histogram) when the parameter values are unknown. In the latter case, different combinations of values must be tried to see which combination yields the lowest chi square score. If there is only one parameter in the theoretical distribution under test, finding an optimum setting of parameter values generally takes relatively few trials because the optimum setting is usually unique and amounts to a "valley" or low spot in the distribution of scores as a function of parameter value. One simply varies the parameter continuously in the direction of lower scores until no further improvement is possible.

The same technique can be used to find optimum fits for a two-parameter distribution such as the logistic-J distribution. However, in both cases, there is a small caveat that must be inserted; the score values of the chi square test are not always continuous functions of the parameter values, owing to the "rule of five," as described in the foregoing section. As parameter values change in the direction of decreasing scores, the chi square value can jump suddenly when the software regroups the categories, causing a discontinuity in score values. However, it can happen that by skipping over the discontinuity, lower scores are achieved once again. But such pockets (local minima) can occasionally bedevil the optimization process. The only surefire method of finding an optimal score is to scan across all possible combinations of parameter values.

**example:** Study # 25 Estuarine Fish in Costa Rica (Winemiller & Mitchell 1992)

In this example, the search for an optimum fit began with the (more or less arbitrary) starting values of $\varepsilon = 1.0$ and $\Delta = 700$. Table 4.3 records progress toward the optimum fit via a score function that was calculated by dividing the chi square score by the number of degrees of freedom. Because different parameter values may produce different numbers of categories and, consequently, different degrees of freedom, one may divide the chi square score by the degrees of freedom to obtain a score that enables one to compare outcomes for fits that involve differing degrees of freedom, especially in the critical neighbourhood of 1.0. Table 4.3 lists these scores, allowing the reader to keep track of progress.

| Iteration | Epsilon | Delta | Score | Chi square |
|---|---|---|---|---|
| 1 | 1.00 | 700 | 1.213 | |
| 2 | 1.02 | 700 | 1.292 | |
| 3 | 0.98 | 700 | 1.208 | |
| 4 | 0.96 | 700 | 1.115 | |
| 5 | 0.95 | 700 | 1.110 | |
| 6 | 0.95 | 690 | 1.109 | |
| 7 | 0.93 | 690 | 1.099 | |
| 8 | 0.91 | 690 | 1.090 | |
| 9 | 0.89 | 690 | 1.080 | |
| 10 | 0.89 | 680 | 1.079 | |
| 11 | 0.87 | 680 | 1.073 | |
| 12 | 0.85 | 680 | 1.063 | 19.155/18 |
| 13 | 0.83 | 680 | 1.100 | |
| 14 | 0.84 | 680 | 1.103 | |
| 15 | 0.85 | 670 | 1.067 | |
| 16 | 0.85 | 675 | 1.067 | |

**Table 4.3** Progress toward a minimum chi square score

As the first row of Table 4.3 indicates, the fitting process produced an initial score of 1.213. Increasing ε to from 1.00 to 1.02 produced a higher score of 1.292, so obviously one wanted to decrease ε, instead of increasing it. Thus a value of 0.98 resulted in a score of 1.208 and subsequent reductions in the value of ε down to 0.85, where the chi square score was 19.155 at 18 degrees of freedom. The fit is a little higher than average, as best fits go. In the last four steps I checked optimality by bracketing the optimal score by higher adjacent ones.

The optimization process runs more smoothly if one alternates between parameters in adjusting their values up and down. It turned out that a decrease in the value of Δ by 10 (row 6) produced a (slightly) lower score, en passant. The small size of the improvement told us that changes in Δ were having little effect on the score, so I returned to ε, with three more decrements taking the score down to 1.080, where another decrease in Δ to 680 again had only a slight effect. Decrementing ε three more times ended with a slight increase in the score, so ε was incremented back to 0.85 and Δ decreased once more. Further changes in either parameter resulted in no improvement over the score already achieved (line 12), so the values of ε and Δ at line 12 were taken as the ones producing a minimum score. Since the response surface is basically bowl-shaped, with the minimum score at the bottom of the bowl, the region in the neighbourhood of the minimum is nearly flat and small changes in either parameter result in little noticeable improvement.

### 4.4 The Kolmogoroff-Smirnov (K-S) test

The Kolmogoroff-Smirnov test (Hays & Winkler 1971) is somewhat simpler than the chi square test, as it does not employ sums of differences. It helps one determine if two sets of data follow the same distribution. Normally the datasets are both derived from the field, but the K-S test may be adapted to curve-fitting, if one of the "datasets" is in fact derived from a theoretical distribution with particular parameter values.

If the two distributions have n categories, one first calculates the corresponding cumulative distributions by adding up, for each value of k, all the entries in the respective distributions up to k. We will denote the respective cumulative values by F(k) (the empirical data) and G(k) (the corresponding theoretical values) respectively. The K-S statistic has a simple formula:

$$D = \max \{|F(k) - G(k)|; k = 1, 2, . . n\}$$

In words, D is simply the maximum absolute difference that occurs over the entire range of abundance categories.

Used in hypothesis testing mode, the results of a K-S test may be compared to a table of critical values, as was the case with the chi square test. If the test score exceeds the critical value for a confidence level of 95%, the hypothesis that the two sets of values arise from the same distribution is rejected. When rejected at this level, one interprets the outcome as follows: "The two sets of data fail to follow the same distribution, with a 5% probability of being wrong.

In a Kolmogoroff-Smirnov table, there is a critical value for each possible size of sample; the larger the sample, the larger the critical value must be. "Acceptance" of the hypothesis has the same interpretation for the Kolmogoroff-Smirnov test as it does for the chi square test; acceptance does not actually imply that the hypothesis is true, as there may be many theoretical candidates that would produce better fits. To confirm the presence of a particular underlying distribution, one needs many (say 50) sets of data, testing each in turn and compiling the results.

We will return to the K-S test in Chapter 8, where it is applied to what might be called "fossil J-curves," namely taxonomic abundance distributions, where one replaces counts of individuals in the present theory by counts of a lower taxon, as distributed across a higher one. For example, in a given geographic region (possibly the entire planet) there might be 21 genera with only one species, 14 with two species, 5 genera with 3 species, and so on. In other words, in this extension of the theory, the lower taxon plays the role of individuals, while the higher taxon plays the role of species.

### 4.5 Application example: sample overlap and similarity

The determination of sample overlap has its uses in field studies of richness, as in the studies cited below. It is statistically impossible to provide an unbiased estimate of overlap without knowing the underlying distribution of species over abundances.

Given two samples of the same community, how many species appear in both samples? For example, in his classic study of the Savannah River, J. Cairns Jr. (1969) used sample overlap to determine the similarity of communities. While the degree of overlap between two unbiased samples must, in a statistical sense, reflect the degree of similarity between the respective communities, the degree of similarity does not have a straightforward interpretation. It turns out that two *identical* communities will typically produce overlaps in the 70 - 80 percent range with typical sample sizes. Thus an overlap of 75 percent, far from indicating a 3/4 overlap, may indicate a near-identity between the respective communities.

The similarity index, as derived in the analysis below, was used by the author (Dewdney 2010) in a study of benthic microbiota in a slow-moving river. The index is based on the empirical shape of the species/abundance distribution in the samples themselves and could therefore be called "parameter-free." The combined samples described in the next paragraph could be replaced by a best-fit logistic-J distribution with some hope of performing even better, but the difference between the two approaches has yet to be tested. In any event, the distribution obtained from the combined samples had the logistic-J shape and it's not clear that the results would be much different.

Suppose one takes two samples $S_1$ and $S_2$ of sizes $N_1$ and $N_2$ (number of individual organisms) from a community and suppose that the ith species appears $a_i$ times in the combined samples. The ratio $a_i/(N_1+N_2)$ then yields an unbiased estimate, $p_i$, of the relative abundance of the ith species in the community. This represents an estimate of the probability that an individual of the ith species will appear if one draws a single organism from the area sampled.

It follows that the probability of this species *not* appearing in such a drawing must be $q_i = (1-p_i)$. Therefore the probability of this species not appearing in a sample of size $N_1$ is $q_i^{N_1}$ and the complementary probability,

$$1- q_i^{N_1},$$

represents the probability that the species will appear at least once in a sample of size $N_1$. Consequently, the probability of the ith species appearing at least once in another sample, this one of size $N_2$, is

$$1- q_i^{N_2},$$

and the probability of the ith species showing up in both samples is simply the product of the two expressions:

$$(1- q_i^{N_2})(1- q_i^{N_1})$$

The expected overlap of the community with itself would then be given by the formula

$$E(S_1, S_2) = \sum(1- q_i^{N_2})(1- q_i^{N_1}), \qquad \text{(i)}$$

where the summation is taken over the union of species in the two samples. Naturally, since the samples are taken from the same community, one would expect the numerical value of formula (i) to be close to the actual overlap.

The same formula may now be applied to the case where the samples $S_1$ and $S_2$ are drawn from different communities, $C_1$ and $C_2$, respectively. In this case, the value of E will reflect what the overlap would be *if* the two communities were the same. In reality, to the extent that the communities are different, the observed overlap will fall below the expected figure, E. Thus the ratio of observed overlap to expected overlap gives a reasonable measure of the degree of real overlap of the communities themselves. Although one is tempted to call the resulting measure the "community overlap," something like that is meant by the term "community similarity" or, more simply, the "similarity index."

We define the *similarity index* for two samples, $S_1$ and $S_2$, by the formula,

$$SI(S_1, S_2) = O(S_1, S_2)/E(S_1, S_2),$$

where $O(S_1, S_2)$ represents the observed overlap between $S_1$ and $S_2$, namely the number of species they have in common. This index represents essentially the true overlap normalized by the expected overlap to give a meaningful, full-scale (0 to 100, when expressed as a percentage) *estimate* of the degree of overlap of the respective communities.

In the study just cited, the following overlaps between samples drawn from different sites were observed. The sites were labeled with a T (transect) code number, as shown in the following tables. Table 4.4 shows the raw overlap figures, the number of species in common between the pairs of samples indicated by the table entry position.

| **Counts** | T5 | T6A | T6B | T7A | T7B |
|---|---|---|---|---|---|
| T4 | 26 | 22 | 28 | 37 | 33 |
| T5 | | 17 | 19 | 25 | 21 |
| T6A | | | 18 | 23 | 19 |
| T6B | | | | 27 | 28 |
| T7A | | | | | 50 |

**Table 4.4.** raw overlap counts between samples

Note that in Table 4.5 the expected numbers of common species are generally higher than the raw overlaps shown in Table 4.4, owing to the fact that the communities are different.

| **Expected** | T5 | T6A | T6B | T7A | T7B |
|---|---|---|---|---|---|
| T4 | 29.9 | 35.9 | 38.8 | 56.0 | 58.9 |
| T5 | | 24.3 | 26.3 | 38.9 | 41.0 |
| T6A | | | 29.1 | 44.4 | 47.6 |
| T6B | | | | 47.6 | 50.3 |
| T7A | | | | | 57.6 |

**Table 4.5.** expected numbers of  species in common

| % | T5 | T6A | T6B | T7A | T7B |
|---|---|---|---|---|---|
| T4 | 87.0 | 61.3 | 72.2 | 67.9 | 56.0 |
| T5 | | 70.0 | 72.2 | 64.3 | 51.2 |
| T6A | | | 61.9 | 51.8 | 39.9 |
| T6B | | | | 56.7 | 55.7 |
| T7A | | | | | 65.8 |

**Table 4.6.** similarity indices for all pairs of samples

In Table 4.6 the similarity indices range from 39.9 to 87.0. It is permissible to interpret these as percentages, as in claiming that T4 and T5 are 87.0 % similar. It is consistent with the high degree of similarity between T4 and T5 that the similarity between either sample and the remaining ones should be relatively close together.

The similarity index technique was not only applied to different communities at the same time, but the same community at different times, yielding a measure of change in the community over the period in question.