Frequency Sampling in Microhistological Studies: An Alternative Model

BYRON K. WILLIAMS

Abstract

Frequency sampling in microhistological studies is discussed in terms of sampling procedures, statistical properties, and biological inferences. Two sampling approaches are described and contrasted, and some standard methods for improving the stability of density estimators are discussed. Possible sources of difficulty are highlighted in terms of sampling design and statistical analysis. An alternative model is proposed that accounts for 2-stage sampling, and yields reasonable, well-behaved estimates of relative densities.

Key Words: sampling methods, digestibility, fecal sampling

Frequency sampling is often used for density estimation in biological studies. A typical application involves the determination of presence or absence of some character within quadrats, followed by a transformation of frequency counts to produce an estimate of density or relative density. Density estimates are based on an application of the frequency-density relationship

$$F = 1 - \exp(-d)\,,$$

a well-known transformation relating character density to quadrat occupancy (Fraker and Brischle 1944, Sparks and Malecheck 1968, Free et al. 1970, Dearden et al. 1975, Holechek and Gross 1982). I focus here on applications of frequency sampling in microhistological studies, in which microscope slides are sampled for presence/absence of plant fragments within microscope fields. A coherent development of frequency sampling for this situation is needed, since most expositions of the frequency-density problem are heuristic (e.g., Grieg-Smith 1964), and leave the important statistical questions unanswered (though see Swindel (1983) for a rigorous analysis of a problem involving frequency-density transformations). In what follows I indicate some sources of difficulty in the use of frequency sampling in microhistological applications and suggest an alternative model for frequency sampling that overcomes these problems.

Sampling Procedures

The general sampling situation typically involves a 2-stage procedure for sampling a mixture of species. The objective is to make inferences about a mixture composed of plant fragments from several distinguishable taxonomic groups. A laboratory process of drying, grinding, and mixing results in a random assortment of fragments, approximately uniform in size and weight. Samples of fragments taken from the mixture are distributed randomly on microscope slides, which then are examined at 100X to 125X. Slide preparation, involving dispersal of a random number of plant fragments, is the first stage of the sampling procedure. The second stage consists of sampling for presence or absence of each species within a series of microscope fields. A count is made of the number of fields containing each species, and these counts (the "frequencies of occurrence") are used to estimate relative densities.

There are at least 2 alternatives for sampling of the microscope slides. The first is similar to methods used in vegetation sampling (Pielou 1976) and involves the division of a slide into, say, k non-overlapping fields. With the slide thus partitioned, each field is examined for presence/absence of plant fragments in each taxon. Since the slide is partitioned into k fields, the size of each field is inversely related to the number of fields by

$$k \times a = A, \tag{1}$$

where A is the area of the slide. Every particle on the slide necessarily falls into 1 of these fields, so that a randomly placed particle will occur in any particular field with probability 1/k. If n_i particles of taxon i are distributed over the slide, the probability that a particular field will be empty of these particles is

$$(1 - 1/k)^{n_i}.$$

A census of the slide then yields the frequency-density transform, based on the total number y_i of fields occupied by particles of the taxon:

$$\begin{aligned} E(y_i) &= k[1 - (1\text{-}1/k)^{n_i}] \\ &\simeq k(1 - e^{-n_i/k})\,, \end{aligned}$$

where $E(y_i)$ is the expected number of fields occupied by taxon i. The average relative frequency F_i is given by

$$\begin{aligned} F_i &= E(y_i/k) \\ &\simeq 1 - e^{-n_i/k} \\ &= 1 - e^{-d_i}\,, \end{aligned} \tag{2}$$

where $d_i = n_i/k$ is defined as the particle density.

This is the sampling procedure and the corresponding transformation described in Johnson (1982). Its distinguishing features are:

- The entire slide is sampled, and therefore every particle is observed.
- "Particle density" is defined in terms of the number of fields that are sampled.
- The number of fields examined is inversely related to the size of the fields, as shown in equation (1).
- The chances of observing particles in any given field decreases with the number of fields examined.

A second approach involves the sampling of a slide (rather than a census of it) by means of non-overlapping fields of area a. A randomly placed particle occurs in any particular field with probability a/A, and the probability that the field will be empty of fragments of taxon i is

$$(1 - a/A)^{n_i}.$$

If the slide is sampled with, say, k such fields, then the expected number y_i of occupied fields is given by

$$E(y_i) = k(1 - (1\text{-}a/A)^{n_i}).$$

This produces an average relative frequency of

$$\begin{aligned} F_i &= (1 - (1\text{-}a/A)^{n_i}) \\ &\simeq 1 - e^{d_i}, \end{aligned} \tag{3}$$

where $d_i = a \times (n_i/A)$ is defined as the particle density.

Key points at which the second approach differs from the first include:

- Only a small fraction of a slide is actually sampled in the second approach. If the slide is observed at 20 or so fields then about 5% to 7% of its surface area is covered. Therefore any given fragment has a high probability of not being observed.
- "Particle density" is *not* defined by the number of fields examined. Rather, it is given in terms of the number of particles per unit area.

Author is with U.S. Fish and Wildlife Service, Patuxent Wildlife Research Center, Laurel, Md. 20708.

Manuscript accepted 1 October 1986.

- There is no relationship between the number of fields to be examined and the size of each field. Within practical limits, decisions about field size and the number of fields examined can be made more or less independently, based on criteria relevant to each decision.

The operational consequences of these differences can be illustrated with a simple example. Assume that particles are randomly distributed over a slide to be investigated with microscope fields. Assume also that a field size is chosen such that an average of 2 particles of taxon i occur within the fields (i.e., $a \times (n_i/A) = 2$). Then sampling according to procedure 2 results in an average relative frequency of about .86, as seen by using $d_i = 2$ in equation (3). Furthermore, this frequency holds irrespective of the number of fields examined. On the other hand, if the slide is partitioned and censused as in procedure 1, the average relative frequency varies with the number of fields. Table 1 displays the results of partitioning the slide into various numbers of fields. Several patterns in the table are worthy of note. First, the average relative frequency changes substantially depending on the number of fields examined. Second, the average relative frequency decreases monotonically from unity as the number of fields increases. Third, the average relative frequency approaches 0 asymptotically as the number of fields increases. It is clear that the "frequency of occurrence" and the corresponding value of "particle density" for this approach are strongly influenced by sampling intensity.

Table 1. Frequencies and densities corresponding to the partioning of a slide into various numbers of fields. F = frequency of occurrence; d = density; k = number of fields.

Number of fields	Procedure 1		Procedure 2	
k	F	d*	F	d
1	1.0	100.	.86	2.
10	.999	10.	.86	2.
20	.99	5.	.86	2.
30	.96	3.3	.86	2.
40	.92	2.5	.86	2.
50	.86	2.	.86	2.
100	.63	1.	.86	2.
200	.39	0.5	.86	2.
400	.22	0.25	.86	2.

*Based on the value n/A = 100.

Given the differences between the 2 procedures, it is somewhat surprising that both procedures result in the same form of a frequency-density relationship. Note, however, that the meaning of "density" (the parameter d_i in equations (2) and (3)) is substantially different for the 2 approaches, so that the meaning of the relationship itself is different. Since the second sampling approach conforms more closely to the microscope procedures used in microhistological studies (a description of sampling procedures in such studies is given by Hansen et al., unpubl.), I base the development below on this approach.

Statistical Properties

Irrespective of the sampling procedure, problems can arise in the analysis of frequency data. These primarily concern the estimation of fragment densities. Typically the number of occupied fields, y_i, is used in equation (2) or (3) to produce estimates of fragment densities, $\hat{d}_i$ (Johnson 1982). Since y_i is random, the estimate y_i/k of $E(y_i/k)$ is also random. Though this estimate is both unbiased and asymptotically efficient (i.e., var (y_i/k) approaches 0 as K increases, the corresponding estimate of d_i is so ill-defined that, regardless of sample size, none of its moments exists. Thus $\hat{d}_i$ has no mean, no variance, no skewness, no kurtosis, nor any moment-based measure of statistical behavior. This is an unfortunate property of $\hat{d}_i$. It means that neither the bias nor the precision of $\hat{d}_i$ can be assessed (or even defined); that standard confidence intervals cannot be specified for $\hat{d}_i$; indeed, that statistical evaluation of $\hat{d}_i$ must necessarily include *ad hoc* adjustments that substantially (and unpredictably) influence the nature of the estimate. In their stated forms the reliability of the density estimates cannot be assessed.

The difficulty with the estimates results from the possible occurrence of 100% occupancy of fields, which corresponds to an infinite value for $\hat{d}_i$. It is sometimes suggested that the problem of 100% occupancy is a minor one that can be handled simply by reducing the frequency of occurrence by some small positive amount δ:

$$F_i = y_i/k - \delta \,.$$

That this proposed adjustment can have important consequences is easily demonstrated by a simple example. Suppose we have a mixture of 2 species, and a sample consisting of 100% occupancy for species 1 and 50% occupancy for species 2. If the value of δ ranges from .1 to .0001, then the relative densities corresponding to species 2 vary from 7% to 22% (Table 2). These results clearly indicate that such adjustments can have potentially important, if largely unanticipated, effects on the biological interpretation of microhistological analysis.

Table 2. Relative densities corresponding to various levels of frequency adjustment, for sample frequencies of 1.0 and .5.

δ	species	$\hat{F}$	$\hat{d}$	$\hat{r}$
.1	1	.9	2.3	.78
	2	.5	.69	.22
.01	1	.99	4.6	.87
	2	.5	.69	.13
.001	1	.999	6.9	.91
	2	.5	.69	.09
.0001	1	.9999	9.2	.93
	2	.5	.69	.07

Two-Stage Sampling

In addition to the sampling and statistical considerations mentioned above, equally important is the issue of the inferential relevance of frequency estimates. To insure that the estimates are biologically relevant, it is necessary to account for the major sources of variation implicit in microhistological studies. There are 2 important sources of variation to be considered. The first is associated with determination of species presence or absence within fields. The second, equally important, source of variation is associated with initial allocation of fragments to a slide. This allocation, which occurs prior to determination of presence or absence, corresponds to the first stage of the sampling procedure described above. Both sources of variation should be included in the statistical model. Unless the initial allocation of fragments is considered, a model essentially focuses on "within-slide" variation, thereby limiting the range of inference to a single slide. Because slide-to-slide variation is not accounted for, statistically valid claims cannot be made about the mixture from which the slides are made.

Thus an alternative formulation of the frequency-density problem is required which accounts for sampling methods and estimation procedures as they are actually practiced. Such a model must include both "among-slide" as well as "within-slide" sources of variation, and it must include sampling restrictions or estimation adjustments to ensure that reasonable, well-behaved density estimators are produced. The following model meets these conditions.

An Alternative Model

The 2-stage nature of the sampling procedure for frequency studies can be accommodated by characterizing each stage statistically. In the first stage we let n be the total number of plant fragments, from all species, that are distributed over a slide:

$$n = \sum_{i=1}^{m} n_i,$$

where m is the number of species. A reasonable distribution for n is the Poisson,

$$g(n) = e^{-\lambda} \lambda^n / n!,$$

with parameter λ (the average number of fragments per slide) under experimental control. Conditional on this total, the distribution of fragments within each species is assumed to be multinomial:

$$f(n_1,\ldots,n_m) = (n!/\prod_{i=1}^{m} n_i!) \prod_{i=1}^{m} r_i^{n_i}$$

where r_i, i=1,...,m are the relative densities. It follows that the unconditional distributions of n_i are independent, and have the form

$$g_i(n_i) = e^{-\lambda_i} (\lambda r_i)^{n_i} / n_i!\,. \qquad (4)$$

These distributions provide a model for the differences in plant fragment densities from slide to slide.

For the second stage of the sampling procedure we model the occupancy count y_i of microscope fields with a binomial distribution

$$f(y_i \mid n_i) = \binom{k}{y_i} p_i^{y_i} (1 - p_i)^{k-y_i}. \qquad (5)$$

The conditional parameter p_i is the probability of occupancy for any particular field, given n_i fragments:

$$p_i = 1 - (1\text{-}a/A)^{n_i},$$

where a is the area encompassed by the microscope field and A is the area of the microscope slide. The distribution shown in equation (5), because it is conditional on fragment density, varies from slide to slide. We seek the unconditional distribution of y_i, since it is y_i on which the frequency-density conversion is based. Its distribution is obtained by combining (4) and (5) into the rather complicated expression

$$f_i(y_i) = e^{-\lambda r_i} \binom{k}{y_i} \sum_{t=0}^{y_i} \binom{y_i}{t} (-1)^t \exp[\lambda r_i (1\text{-}a/A)^{n_i(k-y_i+t)}] \qquad (6)$$

(see the Appendix for a derivation of equation (6)). It is this distribution that relates frequency, y_i/k, and relative density, r_i.

Note that the key addition to the frequency-density model proposed here is reflected in the difference between equation (5) and equation (6). The probabilities p_i in equation (5) are based on a specific number of particles on a given slide. Therefore any inferences to be made apply only to the particular slide under investigation. For the model proposed here, however, the probabilities themselves are assumed to be random, precisely because fragment density n_i varies from slide to slide. Thus equation (6) has the form of a compound distribution (e.g., Pielou 1976), combining components of "slide-to-slide" variation and "within-slide" variation. Estimators based on equation (6) can differ considerably from those based on equation (5).

Unfortunately, a closed form for equation (6) has not been found. Thus it is not possible to simplify the relationship between the observed frequency y_i/k and relative density r_i. Further, since λ and r_i occur together in equation (6), separate estimates of these parameters cannot be obtained for a single species. Note, however, that

$$r_i = \lambda r_i / \sum_{j=1}^{m} (\lambda r_j),$$

so that the usual standardization of the joint estimates of λr_i produces a clean estimate of relative density. Note also that

$$\begin{aligned} E(y_i) &= E_{i_i}[E(y_i \mid n_i)] \\ &= E_{n_i}[k(1 - (1\text{-}a/A)^{n_i})] \\ &\simeq k[1 - \exp(-\lambda a r_i/A)]\,, \end{aligned} \qquad (7)$$

demonstrating that y_i/k is an unbiased estimator of the usual expression relating frequency and density.

As for the statistical properties of density estimators, there are essentially 2 ways to improve their stability. First, a numerical adjustment can be made for density estimates corresponding to high sample frequencies, as indicated earlier. Such adjustments essentially bound the frequency away from 1, where the corresponding value of approaches ∞. For such numerical adjustments, which are currently used by many researchers, there is no obvious way to assess the statistical effect of the adjustment. Thus one cannot unambiguously describe the relationship between the estimate $\hat{r}_i$ and the actual parameter r_i.

The second approach is to use a sampling design that disallows sample frequencies of 1. A reasonable approach is simply to examine additional slides in the event of full occupancy (i.e., if $y_i = k$ for any species). Then the number of slides itself becomes random, with a geometric distribution

$$\text{Prob}(j) = p(1 - p)^j \qquad (8)$$

where j is the number of slides and p is the probability of less than full occupancy for all species. Distributions (6) and (8) can be combined to produce the exact form of a frequency-density relationship (see Appendix), which in turn can be used in the usual estimation procedures. This approach has the advantage that its statistical properties can be described precisely, so that inferences about the species mixture can be supported. Accordingly, it is the procedure recommended here. A formal assessment must consider the probability of full occupancy of all fields based on the distribution in equation (6), and must utilize frequency and density estimators that include frequencies of occurrence across randomly many slides.

Example

The sampling and estimation procedure described above can be illustrated with a simple example using artificial data. Assume that a dietary study is conducted, in which samples of plant fragments are obtained from a fistulated ruminant. Rumen samples are processed and slides prepared according to procedures outlined above. The slides then are examined with 20 fields at 100X for presence or absence of 4 taxonomic groups. For purposes of illustration it is further assumed that the taxonomic assemblage has low equitability, with relative densities given by

$$(r_1, r_4, r_3, r_i) = (.45, .45, .05, .05).$$

To obtain samples in which the rare taxa are represented, it is necessary to prepare slides with high particle densities. As indicated in the Appendix, one effect of high particle density is to increase the probability that all fields are occupied by a single taxon. Thus it may be necessary to sample additional slides.

Now assume that the microscope examination results in frequency counts of (20, 17, 4, 2) for the 4 taxa. Since taxon 1 is represented in all fields, an estimate of density cannot be obtained from equation (2). Thus additional sampling is required. Assume that frequencies of occurrence resulting from examination of a second slide are (17, 19, 1, 2). Since no taxon occupies all the fields on both slides, estimation of relative densities can now proceed. Adding frequency counts across slides and dividing by 40 (the total number of fields examined), we obtain relative frequencies of

$$(\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4) = (.925, .9, .15, .1).$$

Transformation of these values according to equation (2) produces

$$(\hat{d}_1, \hat{d}_2, \hat{d}_3, \hat{d}_4) = (2.6, 2.3, .13, .1),$$

from which are obtained the relative density estimates

$$(\hat{r}_1, \hat{r}_2, \hat{r}_3, \hat{r}_4) = (.5, .45, .03, .02).$$

Several points should be emphasized. First, since estimated relative densities are subject to sampling variability, the correspondence with actual relative densities is unlikely to be exact. For example, if the sampling and estimation procedure were to be replicated a second time, it is highly likely that the estimated relative densities would differ from those in the first replication because of sampling variability alone. The estimates themselves are statistical indicators of relative density, and must be interpreted in light of sampling variability. Indeed, the assessment of density indices against a background of sampling variability should be an integral part of any microhistological study, irrespective of the procedures that are used.

Second, under conditions of taxon inequitability the relative densities of rare species tend to be underestimated. This is reflected in the example by relative density estimates that are below the actual relative density values. Bias occurs because of asymmetry in the frequency-density relationship shown in equation (2). Thus, overestimation of frequencies near 1 has a greater effect on relative density than does underestimation, and the effect is much greater than either overestimation or underestimation of frequencies near zero. The problem of bias can of course be minimized by more intensive sampling, either by examination of more slides or by examination of more fields per slide.

Third, by utilizing averages of frequency counts across slides, the procedure yields estimates of relative density that are statistically improved over those based on a single slide. For example, the estimates are more precise, less biased, and asymptotically normal in their distributions. These properties follow directly from the Central Limit Theorem of classical statistics (e.g., Mood et al. 1974). Were it not for cost considerations, relative density estimation could always be improved by utilizing multiple slides, whether or not full occupancy is a problem.

A comprehensive assessment of the procedure suggested here must address statistical properties of the relative density estimators, based on equation (6), as well as the testing procedures utilizing these estimators. Such an assessment goes well beyond the scope of this paper, and will be reported in a future article.

Literature Cited

Dearden, B.L., R.E. Pegau, and R.M. Hansen. 1975. Precision of microhistological estimates of ruminant food habits. J. Range Manage. 39:402-407.

Free, J.L., R.M. Hansen, and P.L. Sims. 1970. Estimating dry weights of food plants in feces of herbivores. J. Range Manage. 23:300-302.

Fracker, S.B., and J.A. Brischle. 1944. Measuring the local distribution of ribes. Ecology 25:283-303.

Grieg-Smith, P. 1964. Quantitative plant ecology. Butterworth and Co., London.

Hansen, R.M., T.M. Foppe, M.B. Gilbert, R.C. Clark, and H.W. Reynolds. The microhistological analyses of feces as an estimator of herbivore diets. Colorado State Univ., Dep. Range Science. Unpublished report.

Holecheck, J.L., M. Vavra, and R.D. Pieper. 1982. Botanical determination of range herbivore diets: a review. J. Range Manage. 35:309-310.

Johnson, M.L. 1982. Frequency sampling for microscopic analysis of botanical compositions. J. Range Manage. 35:541-542.

Mood, A.M., F.A. Graybill, and D.C. Boes. 1974. Introduction to the theory of statistics. McGraw-Hill, Inc., New York.

Pielou, E.C. 1977. Mathematical ecology. John Wiley and Sons, New York.

Sparks, D.R., and J.C. Malechek. 1968. Estimating percentage dry weight in diets using a microscope technique. J. Range Manage. 21:264-265.

Swindel, B.F. 1983. Choice of size and number of quadrats to estimate density from frequency in Poisson and binomially dispersed populations. Biometrics. 39:455-465.

Appendix

This appendix provides a brief derivation of equation (6), highlights a few of the statistical properties of the resulting density estimators, and indicates how equations (6) and (8) can be combined in a single analysis. From the usual rules of conditional probability (e.g., Mood et al. 1974) we have

$$\begin{aligned}
f_i(y_i) &= \sum_{n_i=0}^{\infty} f(y_i|n_i)\, g_i(n_i) \\
&= \sum_{n_i=0}^{\infty} \binom{k}{y_i} p_i^{y_i}(1-p_i)^{k-y_i} e^{\lambda r_i} (\lambda r_i)^{n_i}/n_i! \\
&= e^{-\lambda r_i} \binom{k}{y_i} \sum_{n_i=0}^{\infty} [1-(1-a/A)^{n_i}]^{y_i} (1-c/A)^{n_i(k-y_i)} (\lambda r_i)/n_i! \\
&= e^{-\lambda r_i} \binom{k}{y_i} = \sum_{n_i=0}^{\infty} \sum_{t=0}^{y_i} \binom{y_i}{t} (-1)^t (1-a/A)^{n_i t} (1-a/A)^{n_i(k-y_i)} (\lambda r_i)^{n_i}/n_i! \\
&= e^{-\lambda r_i} \binom{k}{y_i} \sum_{t=0}^{y_i} \binom{y_i}{t} (-1)^t \sum_{n_i=0}^{\infty} (\lambda r_i)^{n_i} (1-a/A)^{n_i(k-y_i+t)}/n_i! \\
&= e^{-\lambda r_i} \binom{k}{y_i} \sum_{t=0}^{y_i} \binom{y_i}{t} (-1)^t \exp[\lambda r_i(1-a/A^{n_i(k-y_1+t)}]
\end{aligned}$$

Though a closed form for $f_i(y_i)$ has not been found, some of its properties can be ascertained. For example, the expected value of y_i shown in equation (7), increases with

* an increase in λ. Thus the average number of occupied fields increases with the average particle density.
* an increase in r_i. Thus the average number of fields occupied by species i increases with the proportion of species i in the mixture.
* an increase in a/A. Thus the average number of occupied fields increases with field size.

From equation (6) we can express the probability of full occupancy (all fields occupied by species i) by

$$\text{Prob}(n_i=k) = \sum_{t=0}^{k} \binom{k}{t}(-1)^{-t}\exp[-\lambda r_i(1-(1-a/A)^t]$$

If this probability is denoted by q_i, then the probability p in equation (8) of less than full occupancy for all species is given by

$$p = \prod_{j=1}^{s} (1-q_i).$$

Thus the average number of slides required to achieve less than full occupancy for all species is (1-p)/p, which increases in (1-p). But (1-p) in turn increases in q_i. Therefore most slides are required, on average, as

* λ increases. The higher the particle density, the more likely we are to get full occupancy and to require additional slides.
* a/A increases. The larger the field size, the more likely all fields are to be occupied.
* species equitability decreases. The closer we are to the condition that r_i= 4, i=1,...,m, the less likely it is that all fields will be occupied by any one species.

Finally, equations (5) and (7) can be used to assess estimators that utilize random numbers of slides. An example of such an estimator is

$$\hat{F}^i = \sum_{i=1}^{s} y_{ij}/ks$$

where s-1 is the number of slides with full occupancy for at least 1 species and y_i is the frequency count for species i on the jth slide. It can be shown (Williams, in prep.) that this estimator is asymptotically unbiased in p; that is, $E(F_i) \rightarrow F_i$ as $p \rightarrow 1$. Thus the effect of bias is decreased as the average number of slides required goes to 1. Problems arise, however, with mixtures containing species with high relative densities. In that case high particle densities are required in order to pick up the least abundant species, resulting in low values of p. Bias adjustments thus are required for both frequency and relative density estimators.