VOLUME 40, PARTS 3 AND 4 DECEMBER 1953

THE POPULATION FREQUENCIES OF SPECIES AND THE ESTIMATION OF POPULATION PARAMETERS

BY I. J. GOOD

A random sample is drawn from a population of animals of various species. (The theory may also be applied to studies of literary vocabulary, for example.) If a particular species is represented r times in the sample of size N, then r/N is not a good estimate of the population frequency, p, when r is small. Methods are given for estimating p, assuming virtually nothing about the underlying population. The estimates are expressed in terms of smoothed values of the numbers n_r ($r = 1, 2, 3, \ldots$), where n_r is the number of distinct species that are each represented r times in the sample. (n_r may be described as 'the frequency of the frequency r'.) Turing is acknowledged for the most interesting formula in this part of the work. An estimate of the proportion of the population represented by the species occurring in the sample is an immediate corollary. Estimates are made of measures of heterogeneity of the population, including Yule's 'characteristic' and Shannon's 'entropy'. Methods are then discussed that do depend on assumptions about the underlying population. It is here that most work has been done by other writers. It is pointed out that a hypothesis can give a good fit to the numbers n_r but can give quite the wrong value for Yule's characteristic. An example of this is Fisher's fit to some data of Williams's on Macrolepidoptera.

1. *Introduction.* We imagine a random sample to be drawn from an infinite population of animals of various species. Let the sample size be N and let n_r distinct species be each represented exactly r times in the sample, so that

$$\sum_{r=1}^{\infty} r n_r = N. \tag{1}$$

The sample tells us the values of $n_1, n_2, \ldots$, but not of n_0. In fact it is not quite essential that n_0 should be finite though we shall find it convenient to suppose that it is.

We shall suggest a method of estimating, among other things,

(i) the population frequency of each species;

(ii) the total population frequency of all species represented in the sample, or, as we may say, 'the proportion of the population represented by (the species occurring in) the sample';

(iii) various general population parameters measuring heterogeneity, including 'entropy'. By 'general' parameters we mean parameters defined without reference to any special form of hypothesis. In § 7 we shall consider the estimation of parameters for hypotheses of special forms.

Our results are applicable, for example, to studies of literary vocabulary, of accident proneness and of chess openings, but for definiteness we formulate the theory in terms of species of animals.

The formula (2) was first suggested to me, together with an intuitive demonstration, by Dr A. M. Turing several years ago. Hence a very large part of the credit for the present paper should be given to him, and I am most grateful to him for allowing me to publish this work.

Reasonably precise conditions under which our general results are applicable will be given in § 4, but we state at once that the larger is n_1 the more applicable the results. When n_1 is large, n_0 will also be large, but we shall not for the most part attempt to estimate it. There will be a fleeting reference to the estimation of n_0 at the end of § 5 and a few more references in §§ 7 and 8. (See, for example, equation (73).) For populations of known finite size, the problem has been considered by Goodman (1949). He proved that if the sample size is not less than the maximum number of individuals in the population belonging to a single species, then there is only one unbiased estimate of n_0 and he found it. He also pointed out that the unbiased estimate is liable to be unreasonable and suggested some alternative estimates that are always reasonable. There is practically no overlapping between the present work and that of Goodman.

Jeffreys (1948, §3·23) has discussed what is superficially the same problem as (i) above, under the heading 'multiple sampling'. He refers to some earlier work of Johnson (1932). The methods of Johnson and Jeffreys depend on assumptions that, as Jeffreys himself points out, are not always acceptable. Moreover, their methods are not intended to be applicable when n_0 is unknown. The matter is taken up again in §2.

Other work on the frequencies of species has been mainly concerned with the fitting of particular distributions to the data, with or without a theoretical explanation of why these distributions might be expected to be suitable. See, for example, Anscombe (1950), Chambers & Yule (1942), Corbet, Fisher & Williams (1943), Greenwood & Yule (1920), Newbold (1927), Preston (1948), Yule (1944) and Zipf (1932). The methods of the first six sections of the present paper are largely independent of the distributions of population frequencies.

We shall be largely concerned with q_r, the population frequency of an arbitrary species that is represented r times in the sample. We shall use the notation $\mathscr{E}(q_r)$ for the expected value of q_r, in a sense to be explained in §2. Our main result, expressed rather loosely, is that the expected value of q_r is r^*/N, where

$$r^* \simeq (r+1)\, n_{r+1}/n_r. \tag{2}$$

(The symbol '$\simeq$' is used throughout to mean 'is approximately equal to'.) More precisely the n_r's should first be smoothed before applying formula (2). Smoothing is briefly discussed in §3 with examples in §8. If the smoothed values are denoted by $n'_1, n'_2, n'_3, \ldots$, then the more accurate form of equation (2) is

$$r^* \simeq (r+1)\, n'_{r+1}/n'_r. \tag{2$'$}$$

The reader will find it instructive to consider the special case when n'_r is of the Poisson form $s\,e^{-a}a^r/r!$ Then r^* reduces to a constant.

The formula (2) can be generalized to give higher moments of q_r. In fact

$$\mathscr{E}(q_r^m) \simeq \frac{(r+m)^{(m)}}{N^m}\frac{n_{r+m}}{n_r} \quad (r = 1, 2, 3, \ldots;\ m = 0, 1, 2, \ldots), \tag{3}$$

where $t^{(m)} = t(t-1)\ldots(t-m+1)$. We can also write (3) in the form

$$\mathscr{E}(q_r^m) \simeq \mathscr{E}(q_r)\,\mathscr{E}(q_{r+1})\ldots\mathscr{E}(q_{r+m-1}). \tag{4}$$

Moreover, the variance of q_r is

$$V(q_r) \simeq \frac{(r+1)(r+2)}{N^2}\frac{n_{r+2}}{n_r} - \left(\frac{r+1}{N}\frac{n_{r+1}}{n_r}\right)^2$$

$$\simeq \mathscr{E}(q_r)\,[\mathscr{E}(q_{r+1}) - \mathscr{E}(q_r)]. \tag{5}$$

An immediate deduction from (2) is that *the expected total chance of all species that are each represented r times* ($r \geqslant 1$) *in the sample is approximately*

$$(r+1)\, n_{r+1}/N. \tag{6}$$

Hence also the expected total chance of all species that are represented r times or more in the sample is approximately

$$N^{-1}\{(r+1)\,n_{r+1}+(r+2)\,n_{r+2}+\ldots\}. \tag{7}$$

In particular, the expected total chance of all species represented at all in the sample is approximately

$$N^{-1}(2n_2+3n_3+\ldots) = 1-n_1/N. \tag{8}$$

We may say that the proportion of the population represented by the sample is approximately $1-n_1/N$, and *the chance that the next animal sampled will belong to a new species is approximately*

$$n_1/N. \tag{9}$$

(Thus (6) is true even if $r = 0$.)

The results (6), (7), (8) and (9) are improved in accuracy by writing the respective formulae as

$$\frac{(r+1)\,n'_{r+1}n_r}{n'_rN}, \tag{6$'$}$$

$$N^{-1}\left\{\frac{(r+1)\,n'_{r+1}n_r}{n'_r}+\frac{(r+2)\,n'_{r+2}n_{r+1}}{n'_{r+1}}+\ldots\right\}, \tag{7$'$}$$

$$N^{-1}\left\{\frac{2n'_2n_1}{n'_1}+\frac{3n'_3n_2}{n'_2}+\ldots\right\} \tag{8$'$}$$

and

$$\frac{n'_1n_0}{n'_0N}. \tag{9$'$}$$

In most applications this last expression will be extremely close to n'_1/N, and this in its turn will often be very close to n_1/N. It follows that (8′) and (9′) are practically the same as (8) and (9). For the sake of mathematical consistency, the smoothing should be such that (8′) and (9′) add up to 1.

An index of notations used in a fixed sense is given in §9.

I am grateful, and my readers will also be grateful, to Prof. M. G. Kendall for forcing me to clarify some obscurities, especially in §§1 and 2.

2. *Proofs*. Let the number of species in the population be s, which we suppose is finite. This is the same supposition as that n_0 is finite. Our results as far as §6 would be practically unchanged if s were enumerably infinite, but the proofs are more rigorous when it is finite. Let the population frequencies of the species be, in some order, $p_1, p_2, \ldots, p_s$, where

$$p_1+p_2+\ldots+p_s = 1, \quad n_0+n_1+\ldots = s.$$

Let H, or more explicitly $H(p_1, p_2, \ldots, p_s)$, be the statistical hypothesis asserting that $p_1, p_2, \ldots, p_s$ are the population frequencies. We shall discuss the expectation of n_r, given H. It may be objected that the expectation of n_r is simply the observed number n_r, whatever the information, and this objection would be logically correct. Strictly we should introduce extra notation, say $\nu_{r,N}$, for the random variable that is the frequency of the frequency r in a random sample of size N. Then we could introduce the notation $\mathscr{E}(\nu_{r,N} \mid H)$ for the expectation of $\nu_{r,N}$ given H. (Logically this expectation would remain unaffected if particular values of $n_1, n_2, n_3, \ldots$ were given.) In order to avoid the extra notation $\nu_{r,N}$ we shall write $\mathscr{E}(n_r)$ or $\mathscr{E}(n_r \mid H)$ or $\mathscr{E}_N(n_r \mid H)$ instead of $\mathscr{E}(\nu_{r,N} \mid H)$. Confusion can be avoided by reading $\mathscr{E}_N(n_r \mid H)$ as 'the expectation of the frequency of the frequency r when H is given and when the sample size is N'. Similarly, we write $V(n_r) = V(n_r \mid H) = V_N(n_r \mid H)$ for the variance of $\nu_{r,N}$ given H and $\mathscr{E}_N(n_r^2 \mid H)$, etc., for $\mathscr{E}(\nu_{r,N}^2 \mid H)$.

We recall the theorem that an expectation of a sum is the sum of the expectations. It follows that $\mathscr{E}_N(n_r \mid H)$ is the sum over all s species of the probabilities that each will occur r times, given H. So

$$\mathscr{E}_N(n_r \mid H) = \mathscr{E}(n_r \mid H) = \mathscr{E}(n_r)$$
$$= \sum_{\mu=1}^{s} \binom{N}{r} p_\mu^r (1-p_\mu)^{N-r}. \tag{10}$$

In particular

$$\mathscr{E}_N(n_0 \mid H) = \sum_{\mu=1}^{s} (1-p_\mu)^N. \tag{11}$$

If s were infinite this series would diverge. The divergence would be appropriate since n_0 would also be infinite.

Now suppose that in a sample of size N a particular species occurs r times ($r = 0, 1, 2, \ldots$). We shall consider the final (*posterior*) probability that this species is the μth one (of population frequency p_μ). For the sake of rigour it is necessary to define more precisely how the species is selected for consideration. We shall suppose that it is sampled 'at random', or rather equiprobably, from the s species, and that then its number of occurrences in the sample is counted. Thus the initial (*prior*) probability that the species is the μth one is $1/s$. If the species is the μth one then the likelihood that the observed number of occurrences is r is

$$\binom{N}{r} p_\mu^r (1-p_\mu)^{N-r}.$$

We write q_r for the (unknown) population frequency of an arbitrary species that is represented r times in the sample. The final probability that the species is the μth one can be written as $P(q_r = p_\mu \mid H)$ provided that the p_μ's are unequal. (If any of the p_μ's are equal they can be adjusted microscopically so as to be made unequal. These adjustments will have no practical effect.) We may at once deduce the final probability that the species is the μth one by using Bayes's theorem in the form that the final probabilities are proportional to the initial ones times the likelihoods. We find that

$$P\ (q_r = p_\mu \mid H) = \frac{p_\mu^r (1-p_\mu)^{N-r}}{\sum_{\mu=1}^{s} p_\mu^r (1-p_\mu)^{N-r}}. \tag{12}$$

It follows that for any positive integer m,

$$\mathscr{E}(q_r^m \mid H) = \frac{\sum_{\mu=1}^{s} p_\mu^{r+m} (1-p_\mu)^{N-r}}{\sum_{\mu=1}^{s} p_\mu^r (1-p_\mu)^{N-r}} \tag{13}$$

$$= \frac{(r+m)^{(m)}}{(N+m)^{(m)}} \frac{\mathscr{E}_{N+m}(n_{r+m} \mid H)}{\mathscr{E}_N(n_r \mid H)}, \tag{14}$$

in view of (10) and of (10) with N replaced by $N+m$. Immediate consequences of (14) are the basic result

$$\mathscr{E}(q_r \mid H) = \frac{r+1}{N+1} \frac{\mathscr{E}_{N+1}(n_{r+1} \mid H)}{\mathscr{E}_N(n_r \mid H)} \quad (r = 0, 1, 2, \ldots) \tag{15}$$

and

$$V(q_r \mid H) = \frac{\mu'_{r,2,N}\mu'_{r,0,N} - \mu'^2_{r,1,N}}{\mu'^2_{r,0,N}}, \tag{16}$$

where
$$\mu'_{r,t,N} = \frac{(r+t)!}{(N+t)!}\mathscr{E}_{N+t}(n_{r+t} \mid H) \tag{17}$$

$$= \frac{1}{(N-r)!}\sum_{\mu=1}^{s} p_\mu^{r+t}(1-p_\mu)^{N-r} = \frac{r!}{N!}\mathscr{E}_N(n_r \mid H)\,\mathscr{E}(q_r^t \mid H) \tag{18}$$

by (10) and (14). It is clear from either form of (18) that the numbers $\mu'_{r,t,N}$ $(t = 0, 1, 2, \ldots)$ form a sequence of moment constants and therefore satisfy Liapounoff's inequality. (See, for example, Good (1950*a*), or Uspensky (1937).) This checks that the right side of (16) is positive, as it should be being a variance. [It is obvious incidentally that (16) would be true with $\mu'_{r,t,N}$ defined as $\mathscr{E}(q_r^t \mid H)$ times any expression independent of t.]

We can now approximate the formulae (14) and (15) by replacing $\mathscr{E}_{N+m}(n_{r+m} \mid H)$ by the observed value, n_{r+m}, in the sample of size N, or rather by the smoothed value n'_{r+m}. If m is very small compared with N, if n_r and n_{r+m} are not too small and if the sequence $n_1, n_2, n_3, \ldots$ is smoothed in the neighbourhood of n_r and n_{r+m}, then we may expect the approximations to be good. We thus obtain all the approximate results of §1. Note that when the approximation is made of replacing $\mathscr{E}_{N+m}(n_{r+m} \mid H)$ by n'_{r+m} we naturally also change the notation $\mathscr{E}(q_r^m \mid H)$ to $\mathscr{E}(q_r^m)$. For the results become roughly independent of H unless the n_r's are too small to smooth. Observe that $\mathscr{E}(q_r^m \mid H)$ does not depend on the sample, unless H is itself determined by using the sample. On the other hand, $\mathscr{E}(q_r^m)$ does depend on the sample. This may seem a little paradoxical and the following explanation is perhaps worth giving. When we select a particular sequence of smoothed values $n'_1, n'_2, n'_3, \ldots$ we are virtually accepting a particular hypothesis H, say $H\{N;\, n'_1, n'_2, n'_3, \ldots\}$, with curly brackets. (I do not think that this hypothesis is usually a simple statistical hypothesis.) Then $\mathscr{E}(q_r^m)$ can be regarded as a shorthand for $\mathscr{E}(q_r^m \mid H\{N;\, n'_1, n'_2, n'_3, \ldots\})$. (If $H\{\ldots\}$ is not a simple statistical hypothesis this last expression could in theory be given a definite value by assuming a definite distribution of probabilities of the simple statistical hypotheses of which H is a disjunction.) When we regard the smoothing as reasonably reliable we are virtually taking $H\{N;\, n'_1, n'_2, n'_3, \ldots\}$ for granted, as an approximation, so that it can be omitted from the notation without serious risk of confusion. In order to remind ourselves that there is a logical question that is obscured by the notation, we may describe $\mathscr{E}(q_r^m)$ as say a 'credential expectation'.

If a specific H is accepted it is clearly not necessary to use the approximations since equation (13) can then be used directly. Similarly, if H is assumed to be selected from a superpopulation, with an assigned probability density, then again it is theoretically possible to dispense with the approximations. In fact if the 'point' $(p_1, p_2, \ldots, p_s)$ is assumed to be selected from the 'simplex' $p_1 + p_2 + \ldots + p_s = 1$, with probability density proportional to $(p_1 p_2 \ldots p_s)^{k-1}$, where k is a constant, then it is possible to deduce Johnson's estimate $q_r = (r+k)/(N+ks)$. Jeffreys's estimate is the special case $k = 1$, when the probability density is uniform. Jeffreys suggests conditions for the applicability of his estimate, but these conditions are not valid for our problem in general. This is clear if only because we do not assume s to be known.

Jeffreys assumes explicitly that all ordered partitions of N into s non-negative parts are initially equally probable, while Johnson assumes that the probability that the next individual sampled will be of a particular species depends only on N and on the number of times that that species is already represented in the sample. Clearly both methods ignore any information that can be obtained from the entire set of frequencies of all species.

The ignored information is considerable when it is reasonable to smooth the frequencies of the frequencies.

3. *Smoothing.* The purpose of smoothing the sequence $n_1, n_2, n_3, \ldots$ and replacing it by a new sequence $n'_1, n'_2, n'_3, \ldots$, is to be able to make sensible use of the exact results (14) and (15). Ignoring the discrepancy between $\mathscr{E}_N$ and $\mathscr{E}_{N+m}$, the best value of n'_r would be $\mathscr{E}_N(n_r \mid H)$, where H is true. One method of smoothing would be to assume that $H = H(p_1, p_2, \ldots, p_s)$ belongs to some particular set of possible H's, to determine one of these, say H_0, by maximum likelihood and then to calculate n'_r as $\mathscr{E}_N(n_r \mid H_0)$. This method is closely related to that of Fisher in Corbet *et al.* (1943). Since one of our aims is to suggest methods which are virtually distribution-free, it would be most satisfactory to carry out the above method using all possible H's as the set from which to determine H_0. Unfortunately, this theoretically satisfying method leads to a mathematical problem that I have not solved.

It is worth noticing that the sequence $\{\mathscr{E}_N(n_r \mid H)\}$ $(r = 0, 1, 2, \ldots)$ has some properties invariant with respect to changes in H. Ideally the sequence $\{n'_r\}$ should be forced to have these invariant properties. In particular the sequence $\{\mu'_{r,t,N}\}$ $(t = 0, 1, 2, \ldots)$, defined by (17), is a sequence of moment constants. But if $t = o(\sqrt{N})$, then $N^{-t}(r+t)!\, n'_{r+t} \simeq \mu'_{r,t,N}$, so that if $t = o(\sqrt{N})$ we can assume that the sequence $r!\, n'_r$ is a sequence of moment constants and satisfies Liapounoff's inequalities. But this simply implies that $0^*, 1^*, 2^*, \ldots, t^*$ forms an increasing sequence (see equation (2′)), a result which is intuitively obvious even without the restriction $t = o(\sqrt{N})$. (Indeed, the argument could be reversed in order to obtain a new proof of Liapounoff's inequality.) We also intuitively require that $0^*, 1^*, 2^* \ldots$ should itself be a 'smooth' sequence.

Since the sequence $\{\mu'_{r,t,N}\}$ $(t = 0, 1, 2, \ldots)$ is a sequence of moment constants of a probability distribution it follows from Hardy (1949, §11·8) that the sequence is 'totally increasing', i.e. that all its finite differences are non-negative. This result is unfortunately too weak to be useful for our purposes, but it may be possible to make use of some other theorems concerning moment constants. This line of research will not be pursued in the present paper.

A natural principle to adopt when smoothing is that

$$\chi^2 = \sum_{t=1}^{r} \frac{(n'_t - n_t)^2}{V(n_t \mid H)} \tag{19}$$

should not be significant with r degrees of freedom. In §5 we shall obtain an approximate formula for $V(n_r \mid H)$, applicable when $r^2 = o(N)$. The chi-squared test will therefore be applicable when $r^2 = o(N)$. [See formulae (22), (25), (26) and, for particular H's, (65), (85), (86).]

Another similar principle can be understood by thinking of the histogram of n_r as several piles of pennies, n_r pennies in the rth pile. We may visualize the smoothing as the moving of pennies from pile to pile, and we may reasonably insist that pennies moved to the rth pile should not have been moved much further horizontally than a distance $\sqrt{r}$ and almost never further than $2\sqrt{r}$. For $r = 0$ we would not insist on this rule, i.e. we do not insist that $\sum_{r=1}^{\infty} n'_r = \sum_{r=1}^{\infty} n_r$. The analogy with piles of pennies amounts to saying that a species that 'should' have occurred r times is unlikely to have occurred less than $r - \sqrt{r}$ or more than $r + \sqrt{r}$ times.

Let $N' = \Sigma r n'_r$. It seems unnecessary to insist on $N' = N$, provided that N is replaced by N' in such formulae as $\mathscr{E}(q_r) \doteqdot r^*/N$. It will be convenient, however, in § 6 to assume $N' = N$.

For some applications very little smoothing will be required, while for others it may be necessary to use quite elaborate methods. For example, we could

(i) Smooth the n_r's for the range of values of r that interests us, holding in mind the above chi-squared test and the rule concerning $\surd r$. The smoothing techniques may include the use of freehand curves. Rather than working directly with $n_1, n_2, n_3, \ldots$ it may be found more suitable to work with the cumulative sums $n_1, n_1+n_2, n_1+n_2+n_3, \ldots$ or with the cumulative sums of the rn_r or with the logarithms $\log n_1$, $\log n_2$, $\log n_3, \ldots$. There is much to be said for working with the numbers $\surd n_1, \surd n_2, \surd n_3, \ldots$. For if we assume that $V(n_r \mid H)$ is approximately equal to n_r (and in view of (26) and (27) of § 3 this approximation is not on the whole too bad), then it would follow that the standard deviation of $\surd n_r$ is of the order of $\frac{1}{2}$ and therefore largely independent of r. Hence graphical and other smoothing methods can be carried out without having constantly to hold in mind that $\mid n'_r - n_r \mid$ can reasonably take much larger values when n_r is large than when it is small. [The square-root transformation for a Poisson variable, x, was suggested by Bartlett (1936) in order to facilitate the analysis of variance. He showed also that the transformation $\surd(x+\frac{1}{2})$ leads to an even more constant variance. Anscombe (1948) proved that $\surd(x+\frac{3}{8})$ has the most nearly constant variance of any variable of the form $\surd(x+c)$, namely, $\frac{1}{4}$, when the mean of x is large. He attributes this result to A. H. L. Johnson.]

(ii) Calculate $(r+1)\, n'_{r+1}/n'_r$.

(iii) Smooth these values getting, say, r^*.

(iv) Possibly use the values of r^* to improve the smoothing of the n_r's. If this makes a serious difference it will be necessary to check again that the chi-squared test and the $\surd r$ rule have not been violated.

(v) Light can be shed on the reliability of the estimates of the q_r's, etc., if the data are smoothed two or three times, possibly by different people.

In short, the estimation of the q_r's should be done in such a way as to be consistent with the axioms of probability and also with any intuitive judgements that the users of the method are not prepared to abandon or to modify. (This recommendation applies to much more general theoretical scientific work, though there are rare occasions when it may be preferred to abandon the axioms of a science.)

An objection could be raised to the methods of smoothing suggested in the present section. It could be argued that all smoothing methods indirectly assume something about the distribution p_μ, and that one might just as well apply the method of Greenwood & Yule (1920) and its modification by Corbet *et al.* (1943) of assuming a distribution of Pearson's Type III, $Ap^{\alpha} e^{-\beta p}$, or of some other form. Our reply would be that smoothing can be done by making only *local* assumptions, for example, that the square root of $\mathscr{E}(n_r \mid H)$, as a function of r, is approximately 'parabolic' for any nine consecutive values of r. Moreover, it may often be more convenient to apply the general methods of the present section than to attempt to find an adequate hypothesis, H.

4. *Conditions for the applicability of the results of* §§ 1 *and* 2. The condition for the applicability of the results of §§ 1 and 2 is that the user of the methods should be satisfied with his approximations to $\mathscr{E}_{N+m}(n_{r+m} \mid H)$ corresponding to the values of r and m used in the application. This condition is clearly correct, since equation (14) is exact. In particular, if

n_1 is large enough the user would be quite happy to deduce (9) from (15) with $r = 0$. Similarly, he will be satisfied with the estimates of say q_1, q_2 and q_3 provided he is satisfied with the smoothed values (n'_1, n'_2, n'_3, n'_4) of n_1, n_2, n_3 and n_4.

5. *The variance of n_r.* For the application of the chi-squared test described in §3 we need to know more about $V(n_r)$. We begin by obtaining an exact formula for $V(n_r \mid H) = V_N(n_r \mid H)$ and we then make approximations that justify the omission of the symbol H from the notation. It is convenient to introduce the random variable $x_{\mu,r} = x_\mu$ that is defined as 1 if the 'μth species' (of population frequency p_μ) occurs precisely r times in a sample of size N (H being given), otherwise $x_\mu = 0$. Clearly $P\,(x_\mu = 1 \mid H) = \binom{N}{r} p_\mu^r (1-p_\mu)^{N-r}$. Now

$$\begin{aligned}
\mathscr{E}(n_r^2 \mid H) &= \mathscr{E}\big(\sum_\mu x_\mu\big)^2 \\
&= \sum_{\mu,\nu} \mathscr{E}(x_\mu x_\nu) \\
&= \sum_\mu \mathscr{E}(x_\mu) + \sum_{\mu,\nu}^{\mu\neq\nu} \mathscr{E}(x_\mu x_\nu) \\
&= \mathscr{E}(n_r \mid H) + \frac{N!}{r!\,r!\,(N-2r)!} \sum_{\mu,\nu}^{\mu\neq\nu} p_\mu^r p_\nu^r (1-p_\mu-p_\nu)^{N-2r}.
\end{aligned} \tag{20}$$

This is exact. We now make some approximations of the sort used in deriving the Poisson distribution from the binomial. We get, assuming r^2/N, rp_μ and rp_ν to be small,

$$\binom{N}{r} p_\mu^r (1-p_\mu)^{N-r} \simeq \frac{(Np_\mu)^r\, e^{-Np_\mu}}{r!} = a_\mu, \text{ say,}$$

and

$$\frac{N!}{r!\,r!\,(N-2r)!} p_\mu^r p_\nu^r (1-p_\mu-p_\nu)^{N-2r} \simeq a_\mu a_\nu.$$

Moreover, it is intuitively clear that terms for which p_μ or p_ν is far from r/N can make no serious contribution to the summation in (20). Hence, if $r^2 = o(N)$,

$$\begin{aligned}
\mathscr{E}(n_r^2 \mid H) &\simeq \mathscr{E}(n_r \mid H) + \sum_{\mu,\nu}^{\mu\neq\nu} a_\mu a_\nu \\
&\simeq \mathscr{E}(n_r \mid H) + \{\mathscr{E}(n_r \mid H)\}^2 - \sum_\mu a_\mu^2.
\end{aligned}$$

Therefore the variance of n_r for samples of size N is

$$V_N(n_r \mid H) \simeq \mathscr{E}_N(n_r \mid H) - \sum_\mu \frac{(Np_\mu)^{2r}\, e^{-2Np_\mu}}{(r!)^2} \tag{21}$$

$$\simeq \mathscr{E}_N(n_r \mid H) - \frac{1}{2^{2r}} \binom{2r}{r} \mathscr{E}_{2N}(n_{2r} \mid H). \tag{22}$$

Formulae (21) and (22) are elegant but need further transformation, when H is unknown, before they can be used for calculation. Notice first that there are n_u species whose expected population frequencies are q_u $(u = 0, 1, 2, \ldots)$. Hence we have for $r = 0, 1, 2, \ldots$; $r^2 = o(N)$,

$$\begin{aligned}
V(n_r \mid H) &\simeq \mathscr{E}(n_r \mid H) - \sum_{u=0}^{\infty} n_u \left\{\frac{(Nq_u)^r\, e^{-Nq_u}}{r!}\right\}^2 \\
&= \mathscr{E}(n_r \mid H) - \sum_{u=0}^{\infty} n_u \left\{\frac{u^{*r}\, e^{-u^*}}{r!}\right\}^2.
\end{aligned} \tag{23}$$

Similarly and rather more simply, when $r^2 = o(N)$,

$$\mathscr{E}(n_r \mid H) \overset{\frown}{=} \frac{1}{r!} \sum_{u=0}^{\infty} n_u u^{*r} e^{-u^*}. \tag{24}$$

Now for any positive x, $x^r e^{-x} \leqslant r^r e^{-r}$, so

$$\mathscr{E}(n_r \mid H) \overset{\frown}{\geqq} V(n_r \mid H) \overset{\frown}{\geqq} \mathscr{E}(n_r \mid H)\left(1 - \frac{r^r e^{-r}}{r!}\right). \tag{25}$$

Using Stirling's formula for $r-1$ we have

$$\mathscr{E}(n_r \mid H) \overset{\frown}{\geqq} V(n_r \mid H) \overset{\frown}{\geqq} \mathscr{E}(n_r \mid H)\left(1 - \frac{1}{\sqrt{(2\pi r)}}\right) \quad (r = 2, 3, \ldots), \tag{26}$$

while

$$\mathscr{E}(n_1 \mid H) \overset{\frown}{\geqq} V(n_1 \mid H) \overset{\frown}{\geqq} \mathscr{E}(n_1 \mid H)(1 - e^{-1}) \tag{27}$$

(see also formula (65) in §7). Now the most desirable value for n'_r would be $\mathscr{E}(n_r \mid H)$ where H is true, so if our smoothing of the n_r's is to be satisfactory for any particular values of r small compared with $\sqrt{N}$ we may write

$$n'_r \overset{\frown}{=} \sum_{u=0}^{\infty} n_u \frac{u^{*r} e^{-u^*}}{r!}, \tag{28}$$

and these approximate equations may be used as a test of consistency for the values of n'_r and u^*. Indeed, it may be possible iteratively to solve equations (28) combined with (2′) and thus very systematically to obtain estimates of n'_r and r^* for values of r small compared with $\sqrt{N}$. This iterative process may possibly lead to estimates of n'_0 and 0^*, but I have not yet tried out the process. For most applications the less systematic methods previously described will probably prove to be adequate, and any smoothing obtained by these methods can be partially tested by means of χ^2 in the form (19), together with the inequalities (26) and (27). (See also the remarks following equations (65) and (87).)

6. *Estimation of some population parameters, including entropy.* Let us consider the population parameters

$$c_{m,n} = \sum_{\mu=1}^{s} p_\mu^m (-\log p_\mu)^n \quad (m, n = 0, 1, 2, \ldots), \tag{29}$$

which can be regarded as measures of heterogeneity of the population. The sequence $c_{0,0} = 1, c_{1,0} = s, c_{2,0}, c_{3,0}, \ldots$ may be called the 'moment constants' of the population, while $c_{1\,1}$ is called the 'entropy' in the modern theory of communication (see Shannon, 1948). More generally, $c_{1,n}$ is the moment about zero of the amount of information from each selection of an animal (or word), where 'amount of information' is here used in the sense of Good (1950*b*, p. 75), i.e. as minus the logarithm of a probability. (The last sentence of p. 75 of this reference is incorrect, as Prof. M. S. Bartlett has pointed out.) We find it no more difficult to give estimates of $c_{m,n}$ than of $c_{1,n}$, at any rate when $n = 0$ or 1.

It is an immediate consequence of (10) that an unbiased estimate of $c_{m,0}$ is

$$\hat{c}_{m,0} = \frac{1}{N^{(m)}} \sum_r r^{(m)} n_r. \tag{30}$$

$\hat{c}_{2,0}$ is in effect used by Yule (1944) to measure the heterogeneity of samples of vocabulary, and he calls $10{,}000\hat{c}_{2,0}(1-1/N)$ the 'characteristic' of the material. The sequence of all sampling moments of $\hat{c}_{2,0}$ involves all the population parameters $c_{m,0}$. For example, as pointed out by Simpson (1949), for large N,

$$V(\hat{c}_{2,0}) \overset{\frown}{=} \frac{4}{N}(c_{3,0} - c_{2,0}^2). \tag{30 A}$$

Unbiased statistics are rather unfashionable nowadays, partly because they can take impossible values. For example, $\hat{c}_{m,0}$ could vanish, although it is easy to see that $c_{m,0} \geqslant s^{-(m-1)}$. (Compare Good (1950*b*, p. 103), where estimates of $c_{m,0}$ are implicit for general multinomial distributions, no attempt being made to smooth the n_r's.) We shall find estimates of $c_{m,1}$ and also estimates of $c_{m,0}$ that are at least sometimes better than $\hat{c}_{m,0}$.

We have

$$c_{m,0} = \frac{1}{N^{(m)}} \sum_r r^{(m)} \mathscr{E}(n_r \mid H), \tag{31}$$

since this is in effect what is meant by saying that $\hat{c}_{m,0}$ is an unbiased estimate of $c_{m,0}$. If the statistician is satisfied with his smoothing, i.e. if he assumes that $n'_r \simeq \mathscr{E}(n_r \mid H)$, and if he has forced $N' = N$, then he can estimate $c_{m,0}$ as

$$\tilde{c}_{m,0} = \frac{1}{N^{(m)}} \sum_r r^{(m)} n'_r, \tag{32}$$

and he will be prepared to assume that this is a more efficient estimate than $\hat{c}_{m,0}$. More generally if the smoothing is satisfactory for $r = 1, 2, \ldots, t$ but not for all larger values of r, then a good estimate of $c_{m,0}$ will be $\tilde{c}_{m,0}(t)$, where

$$\tilde{c}_{m,0}(t) = \frac{1}{N^{(m)}} \left\{ \sum_{r=m}^{t} r^{(m)} n'_r + \sum_{r=t+1}^{\infty} r^{(m)} n_r \right\}. \tag{33}$$

We shall next consider estimates $\tilde{c}_{m,1}$ of $c_{m,1}$. We shall begin by proving that (exactly)

$$c_{m,1} = \frac{1}{N^{(m)}} \sum_r r^{(m)} \mathscr{E}(n_r \mid H) \left\{ \frac{1}{r+1} + \frac{1}{r+2} + \ldots + \frac{1}{N-r} - \frac{d}{dr} \log \mathscr{E}(n_r \mid H) - \mathscr{E}[\log(1-q_r) \mid H] \right\}. \tag{34}$$

The differential coefficient in this expression is made meaningful by means of a suitable definition of $\mathscr{E}(n_r \mid H)$ for non-integral values of r. This definition is obtained from equation (10) by writing $\Gamma(N+1)/\Gamma(r+1)\,\Gamma(N-r+1)$ instead of $\binom{N}{r}$.

In order to prove (34) we shall need the following generalization of (13), valid for any function $f(.)$:

$$\mathscr{E}(n_r \mid H)\, \mathscr{E}[f(q_r) \mid H] = \binom{N}{r} \sum_\mu p_\mu^r (1-p_\mu)^{N-r} f(p_\mu). \tag{35}$$

We also require the following property of the gamma function. If b is a non-negative integer,

$$\frac{\Gamma'(b+1)}{\Gamma(b+1)} = 1 + \frac{1}{2} + \ldots + \frac{1}{b} - \gamma, \tag{36}$$

where $\gamma = 0{\cdot}577215\ldots$ is the Euler-Mascheroni constant. (See, for example, Jeffreys & Jeffreys (1946, §15·04).) It follows from (10) and (36) that

$$\begin{aligned} \frac{d}{dr} \mathscr{E}(n_r \mid H) &= \binom{N}{r} \sum_\mu p_\mu^r (1-p_\mu)^{N-r} \left(\frac{1}{r+1} + \frac{1}{r+2} + \ldots + \frac{1}{N-r} + \log \frac{p_\mu}{1-p_\mu} \right) \\ &= \mathscr{E}(n_r \mid H) \left\{ \frac{1}{r+1} + \ldots + \frac{1}{N-r} + \mathscr{E}\left(\log \frac{q_r}{1-q_r} \mid H \right) \right\}, \end{aligned}$$

by (35). Therefore

$$\mathscr{E}(n_r \mid H)\left(\frac{1}{r+1}+\ldots+\frac{1}{N-r}\right)-\frac{d}{dr}\mathscr{E}(n_r \mid H)-\mathscr{E}(n_r \mid H)\,\mathscr{E}[\log(1-q_r) \mid H]$$
$$=\mathscr{E}(n_r \mid H)\,\mathscr{E}(-\log q_r \mid H)=\binom{N}{r}\sum_\mu p_\mu^r(1-p_\mu)^{N-r}(-\log p_\mu),$$

by (35) again. Multiplying by $r^{(m)}/N^{(m)}$ and summing with respect to r, we find that the right-hand side of (34) equals

$$\sum_\mu\sum_r\binom{N-m}{r-m}p_\mu^{r-m}(1-p_\mu)^{N-m-(r-m)}p_\mu^m(-\log p_\mu)=c_{m,1}$$

as asserted.

$c_{m,2}$ can be evaluated in a similar manner by first writing down $\left(\frac{d}{dr}\right)^2\mathscr{E}(n_r \mid H)$, but the result is complicated and will be omitted.

As in the estimation of $c_{m,0}$, if the statistician is satisfied with his smoothing, then he can write

$$c_{m,1}\doteqdot\frac{1}{N^{(m)}}\sum_r r^{(m)}n'_r\left\{\frac{1}{r+1}+\frac{1}{r+2}+\ldots+\frac{1}{N-r}-\frac{d}{dr}\log n'_r-\mathscr{E}[\log(1-q_r) \mid H]\right\}.$$

If N is large the approximation can be written

$$c_{m,1}\doteqdot\frac{1}{N^{(m)}}\sum_r r^{(m)}n'_r\left\{\log N-\left(1+\frac{1}{2}+\ldots+\frac{1}{r}-\gamma\right)-\frac{d}{dr}\log n'_r-\mathscr{E}\left(\log\frac{1-q_r}{1-\frac{r}{N}} \;\middle|\; H\right)\right\}.$$

Now it is intuitively clear that $\mathscr{E}\left(q_r-\frac{r}{N}\right)$, which equals $\frac{r^*-r}{N}$, must be $O\left(\frac{\sqrt{r}}{N}\right)=o(N)$, and therefore

$$c_{m,1}\doteqdot\frac{1}{N^{(m)}}\sum_r r^{(m)}n'_r\left\{\log N-\left(1+\frac{1}{2}+\ldots+\frac{1}{r}-\gamma\right)-\frac{d}{dr}\log n'_r\right\}$$
$$=\tilde{c}_{m,0}\log N-\frac{1}{N^{(m)}}\sum_r r^{(m)}n'_r\left(g_r+\frac{d}{dr}\log n'_r\right), \tag{37}$$

where

$$g_r=1+\frac{1}{2}+\ldots+\frac{1}{r}-\gamma. \tag{38}$$

In particular, the entropy $c_{1,1}\doteqdot\tilde{c}_{1,1}$, where

$$\tilde{c}_{1,1}=\log N-\frac{1}{N}\sum_r rn'_r\left(g_r+\frac{d}{dr}\log n'_r\right). \tag{39}$$

The differentiation can be performed graphically for all r or by numerical differentiation for $r=3,4,5,\ldots$. (For numerical differentiation see, for example, Jeffreys & Jeffreys (1946, § 9·07).) Another estimate of the entropy is $\hat{\tilde{c}}_{1,1}$, where

$$\hat{\tilde{c}}_{1,1}=\log N-\frac{1}{N}\sum_r rn_r\left(g_r+\frac{d}{dr}\log n'_r\right), \tag{40}$$

in which the 'prime' has been omitted from the first occurrence of n'_r in (39). This estimate, $\hat{\tilde{c}}_{1,1}$, has leanings towards being an unbiased estimate of the entropy. It can hardly be as good as (39) when the smoothing is reliable. Perhaps the best method of using the present theory for estimating $c_{m,1}$ is to use the compromise $\tilde{c}_{m,1}(t)$ defined in the obvious way by analogy with (33). For large values of r, the factor $g_r + \frac{d}{dr}\log n'_r$ may be replaced by $\log r$ to a good approximation. Terms of $\tilde{c}_{m,1}(t)$ for which this approximation is made, i.e. terms of the form $rn_r \log r$ may be regarded as crude and unadjusted.

7. *Special hypotheses, H.* In this section we shall consider some special classes of hypotheses, H, which determine the distribution p_μ. So far we have taken this distribution as discrete for the sake of logical simplicity. In the present section we shall find it convenient to assume that there is a density function, $f(p)$, where $f(p)\,dp$ is the number of species whose population frequencies lie between p and $p+dp$. (The formulae may of course be generalized to arbitrary distributions by using the Stieltjes integral.) Clearly

$$\int_0^1 f(p)\,dp = s, \tag{41}$$

$$\int_0^1 pf(p)\,dp = 1. \tag{42}$$

The expected value of p for an animal at random from the population is

$$\mathscr{E}(p \mid H) = \int_0^1 p^2 f(p)\,dp = c_{2,0}. \tag{43}$$

The appropriate modifications of the previous formulae are obvious. For example, instead of (10) and (20) we have

$$\mathscr{E}_N(n_r \mid H) = \binom{N}{r}\int_0^1 p^r(1-p)^{N-r} f(p)\,dp, \tag{44}$$

$$\begin{aligned}\mathscr{E}_N(n_r^2 \mid H) = \mathscr{E}_N(n_r \mid H) + \binom{N}{r,r}\int_0^1\int_0^1 p^r q^r(1-p-q)^{N-2r} f(p) f(q)\,dp\,dq \\ - \binom{N}{r,r}\int_0^1 p^{2r}(1-2p)^{N-2r} f(p)\,dp.\end{aligned} \tag{45}$$

Notice the elegant checks of (44) and (45) that $\mathscr{E}_0(n_0 \mid H) = s$, $\mathscr{E}_1(n_1 \mid H) = 1$, $V_0(n_0 \mid H) = 0$, $V_1(n_1 \mid H) = 0$. Formula (44) leads to the less precise but often more convenient formula

$$\begin{aligned}\mathscr{E}_N(n_r \mid H) &= \frac{1}{r!}\left[1 + O\left(\frac{r^2}{N}\right)\right]\int_0^1 (pN)^r e^{-pN} f(p)\,dp \\ &\simeq \frac{1}{r!}\int_0^1 (pN)^r e^{-pN} f(p)\,dp \quad (r^2 = o(N)),\end{aligned} \tag{46}$$

while a similar treatment of formula (45) leads back merely to formula (22).

We shall now list a number of different types of possible hypotheses and then discuss them. The normalizing constants are all deduced from (42).

H_1 (Pearson's Type I):

$$f(p) = \frac{(\alpha+\beta+2)!}{(\alpha+1)!\,\beta!}\,p^\alpha(1-p)^\beta \quad (\alpha > -1, \beta > -1). \tag{47}$$

H_2 (Pearson's Type III):

$$f(p) = \frac{\beta^{\alpha+2}}{(\alpha+1)!}\,p^\alpha e^{-\beta p} \quad (\alpha > -1, \beta > 0). \tag{48}$$

H_3 (same as H_2 but with $\alpha = -1$):

$$f(p) = \beta p^{-1} e^{-\beta p} \quad (\beta > 0). \tag{49}$$

H_4:
$$f(p) = sAp^{\alpha} \exp(-\beta p - \epsilon p^{-1}) \quad (\beta > 0, \epsilon > 0). \tag{50}$$

H_5 (truncated form of H_3):
$$f(p) = \begin{cases} \beta p^{-1} e^{-\beta p} & (p > p_0), \\ 0 & (p < p_0). \end{cases} \tag{51}$$

H_6 (truncated form of another special case of H_2):

$$f(p) = \begin{cases} \dfrac{p^{-2} e^{-\beta p}}{E(p_0 \beta)} & (p > p_0), \\ 0 & (p < p_0), \end{cases} \tag{52}$$

where $E(w) = -\text{Ei}(-w) = \int_w^{\infty} u^{-1} e^{-u} du$. Ei (w) is known as the 'exponential integral' and has been tabulated several times. (For a list of these tables see Fletcher, Miller & Rosenhead (1946, §§13·2 and 13·21).)

We list also a few less completely formulated hypotheses, H_7, H_8 and H_9, for which the population is not explicitly specified, but only the values of $\mathscr{E}_N(n_r \mid H)$. Hence for these hypotheses the parameters may depend on N.

H_7 (Zipf laws):
$$\mathscr{E}(n_r \mid H_7) \propto r^{-\zeta} \quad (r \geqslant 1, \zeta > 0), \tag{53}$$

where ζ is often taken as 2 by Zipf. (See also (94) below.)

H_8 (H_7 with a convergence factor):

$$\mathscr{E}(n_r \mid H_8) = \frac{\lambda x^r}{r^{\zeta}} \quad (r \geqslant 1, \zeta > 0, 0 < x < 1). \tag{54}$$

H_9 (a modification of a special case of H_8):

$$\mathscr{E}(n_r \mid H_9) = \frac{\lambda x^r}{r(r+1)} \quad (r \geqslant 1). \tag{55}$$

We now discuss the nine hypotheses.

(i) H_1 has the advantage that the exact formula (44) can be evaluated in elementary terms. We can see from (41) and (43) that

$$s = \frac{\alpha + \beta + 2}{\alpha + 1}, \tag{56}$$

$$\mathscr{E}(p \mid H_1) = \frac{\alpha + 2}{\alpha + \beta + 3} = \frac{(\alpha + 2)(\alpha + \beta + 2)}{(\alpha + 1)(\alpha + \beta + 3)} \frac{1}{s}. \tag{57}$$

In most applications we want $f(p)$ to be small when p is not small and $\mathscr{E}(p \mid H)$ to be large compared with $1/s$. Hence if a hypothesis of the form H_1 is to be appropriate at all, we shall usually want β to be large, by (47), and α to be close to -1, by (57).

By (44) we see that

$$\mathscr{E}_N(n_r \mid H_1) = \binom{N}{r} \frac{(\alpha + \beta + 2)!\,(\alpha + r)!\,(\beta + N - r)!}{(\alpha + 1)!\,\beta!\,(\alpha + \beta + N + 1)!}. \tag{58}$$

Hence, by (2′), if the smoothed values n'_r and n'_{r+1} were equal to their expectations, given H_1, we would have

$$r^* = \frac{(\alpha + r + 1)(N - r)}{\beta + N - r}. \tag{59}$$

(ii) H_2 can be regarded as a convenient approximation to H_1 if $\beta > 0$. Strictly, the hypothesis H_2 is impossible since it allows values of p greater than 1, but it gives all such values of p combined a very small probability provided that β is large. H_2 was used by Green wood & Yule (1920) and by Fisher (see Corbet *et al.* 1943). We have

$$s = \frac{\beta}{\alpha+1}, \quad \mathscr{E}(p \mid H_2) = \frac{\alpha+2}{\beta} = \frac{\alpha+2}{\alpha+1}\frac{1}{s}, \tag{60}$$

so that α must be close to -1. Hence, if $r^2 = o(N)$,

$$\mathscr{E}_N(n_r \mid H_2) \simeq \frac{\beta(\alpha+r)!}{(\alpha+1)!\,r!}\left(\frac{N}{N+\beta}\right)^r \left(\frac{\beta}{N}\right)^{\alpha+1}, \tag{61}$$

which is of the negative binomial form.

(iii) Of all hypotheses of the form H_2, Fisher (Corbet *et al.* 1943) was mainly concerned with H_3, the case $\alpha = -1$. (See example (i) in § 8 below.) Then

$$s = \infty, \quad \mathscr{E}(p \mid H_3) = \frac{1}{\beta}, \tag{62}$$

$$\mathscr{E}_N(n_r \mid H_3) \simeq \frac{\beta}{r}\left(\frac{N}{N+\beta}\right)^r = \frac{\beta x^r}{r} \quad (r^2 = o(N)), \tag{63}$$

say. For large samples, x (which, unlike β, depends on N) is close to 1 and the factor x^r may be regarded as a convergence factor which prevents $\sum\limits_{r=1}^{\infty} \mathscr{E}_N(n_r \mid H_3)$ from becoming infinite. The convergence factor also increases the likelihood of being able to find a satisfactory fit to given frequencies, n_r, merely because it involves a new parameter.

We see from (22) that

$$V_N(n_r \mid H_3) \simeq \mathscr{E}_N(n_r \mid H_3)\left\{1 - \frac{1}{2^{2r+1}}\binom{2r}{r}\left(\frac{N}{N+\beta}\right)^r\right\}. \tag{64}$$

If $\beta r = o(N)$ it follows that

$$\frac{V_N(n_r \mid H_3)}{\mathscr{E}_N(n_r \mid H_3)} \simeq 1 - \frac{1}{2^{2r+1}}\binom{2r}{r}. \tag{65}$$

Thus in these circumstances $V_N(n_r \mid H_3)$ lies between the bounds given by (26) and (27), being for each r about twice as close to the smaller bound than to the larger one. When applying the chi-squared test, where χ^2 is defined by equation (19), we can hardly go far wrong by assuming (65) to be applicable whatever the distribution determined by H may be. But, of course, we may often be able to improve on (65) when H is specified in terms of the distribution of p. For convenience in applying (65) we give a short table of values of

$$k_r = \left[1 - \frac{1}{2^{r+1}}\binom{2r}{r}\right]^{-1} \tag{66}$$

r	k_r	r	k_r	r	k_r
1	1·33	6	1·13	11	1·10
2	1·23	7	1·12	12	1·09
3	1·19	8	1·11	13	1·09
4	1·16	9	1·11	14	1·08
5	1·14	10	1·10	15	1·08

For larger values of r, the approximation $1 + 1/\{(2\sqrt{(\pi r)})\}$ is correct to two places of decimals.

Suppose we are given a sample of size N and we wish to estimate β and x. The method used by Fisher was to equate the observed values of $\Sigma rn_r = N$ and $\Sigma n_r = S$ to their expected values. (Note that S is the observed number of species and should not be confused with s.) This led him to the equations

$$-\beta \log_e(1-x) = S, \quad N = \beta x/(1-x), \tag{67}$$

$$\frac{N}{S} = \frac{x}{-(1-x)\log_e(1-x)}, \tag{68}$$

which he solved by using a table of $x/(1-x)$ in terms of $\log_{10}(N/S)$.

A theoretically more satisfactory method of estimating β and x would be by minimizing χ^2, defined by (19), with $r = \infty$. This method leads to equations which would be most laborious to solve by hand but which will be given here since large-scale computers now exist. To prevent misunderstanding we mention at once that Fisher obtained a perfectly good fit by the simpler method, in his example, i.e. example (i) of §8 below, though, as pointed out in §8, H_3 must not be too literally regarded as true.

By (65) we may write

$$\chi^2 = \sum_{r=1}^{\infty} k_r \left(\frac{\beta x^r}{r} - 2n_r + \frac{rn_r^2}{\beta x^r}\right). \tag{69}$$

The equations giving β and x will then be

$$\beta^2 \Sigma k_r x^r/r = \Sigma r k_r n_r^2 x^{-r}, \tag{70}$$

$$\beta^2 \Sigma k_r x^r = \Sigma r^2 k_r n_r x^{-r}, \tag{71}$$

and these equations could be solved iteratively.

When β and x are specified the cumulative sums of $\mathscr{E}_N(n_r \mid H_3)$ can be found by making use of the approximation

$$\sum_t^{t \geqslant r} \frac{x^t}{t} \bumpeq E(-r\log_e x) + \frac{x^r}{2r}\left(1 + \tfrac{1}{6}\log_e x - \frac{1}{6r}\right), \tag{72}$$

which will be a very good approximation if the terms involving $\frac{1}{6}\log x$ and $\dfrac{1}{6r}$ are negligible. This approximation can be obtained by means of the Euler-Maclaurin summation formula. (See, for example, Whittaker & Watson (1935, §7·21).)

(iv) We have just seen that when $\alpha = -1$ in H_2 we obtain $s = \infty$ and of course $\mathscr{E}_N(n_0 \mid H) = \infty$. There are strong indications in examples (ii), (iii) and (iv) of §8 that we may wish to take $\alpha < -1$, and then even worse divergencies occur. For example, if $\alpha = -2$ we would obtain, from (61), the intolerable result

$$\mathscr{E}_N(n_1 \mid H_2)/\mathscr{E}_N(n_2 \mid H_2) = \infty.$$

In order to avoid these divergencies we could in theory use hypothesis H_4, with a small value of ϵ. Unfortunately, this hypothesis seems to be analytically unwieldy; it is mentioned partly for its interest as intermediate between Pearson's Types III and V.

(v) Another method of avoiding divergencies is to use truncated distributions. These truncated distributions are not theoretically pleasing but at least some of them can be handled analytically. H_5 is a truncated form of H_3. We may describe p_0 as the smallest possible population frequency of any species. In most applications it would be difficult to obtain a sample large enough to determine p_0 with any accuracy. In fact if the estimate of

p_0 were to be reliable the sample would need to be so large that n_r would vanish for all small values of r. In the examples of § 8, n_1 is always larger than any other value of n_r, so these samples would need to be increased greatly before one could expect even n_1 to vanish.

We obtain from (41)

$$s = \beta E(p_0 \beta). \tag{73}$$

Now

$$E(w) = -\gamma - \log_e w + w - \frac{w^2}{2!\,2} + \frac{w^3}{3!\,3} - \ldots, \tag{74}$$

an equation which is undoubtedly well known. It can be proved, for example, by using Dirichlet's formula for γ. (See, for example, Whittaker & Watson (1935, § 12·3, example 2).) In particular, if w is very small,

$$E(w) \sim -\log_e (\gamma' w), \tag{75}$$

where

$$\gamma' = e^{\gamma} = 1{\cdot}781072. \tag{76}$$

(Cf. Jahnke & Emde (1933, p. 79), where our γ' is denoted by γ.) Since p_0 is assumed to be small, we have

$$s \bumpeq -\beta \log (p_0 \gamma' \beta), \quad p_0 \bumpeq \beta^{-1} e^{-\gamma - s/\beta}. \tag{77}$$

On applying equation (46) we see that

$$\mathscr{E}_N(n_r \mid H_5) \bumpeq \frac{\beta}{r} \left(\frac{N}{N+\beta} \right)^r \quad (r \geqslant 1, r^2 = o(N)), \tag{78}$$

$$\mathscr{E}_N(n_0 \mid H_5) \bumpeq \beta E[p_0(N+\beta)] \bumpeq -\beta \log [\beta_0 \gamma' (N+\beta)]. \tag{79}$$

The check may be noticed that equations (77), (78) and (79) are consistent with

$$s = \mathscr{E}(s \mid H_5) = \sum_{r=0}^{\infty} \mathscr{E}_N(n_r \mid H_5).$$

Formula (77) is of some interest, but in most applications both p_0 and s will be largely metaphysical, i.e. observable only within very wide proportional limits.

(vi) The difficulty of determining p_0 would not apply to the same extent if $\alpha = -2$, i.e. for hypothesis H_6. (This hypothesis is fairly appropriate for example (iv) of § 8.) We have, by (46),

$$\mathscr{E}_N(n_1 \mid H_6) \bumpeq \lambda x E[p_0(\beta + N)] \bumpeq -\lambda x \log \frac{p_0 \gamma' N}{x}, \tag{80}$$

$$\mathscr{E}_N(n_r \mid H_6) \bumpeq \frac{\lambda x^r}{r(r-1)} \quad (r \geqslant 2, r^2 = o(N)), \tag{81}$$

where x and λ, unlike β and p_0, depend on N and are given by

$$x = \frac{N}{N+\beta} \tag{82}$$

and

$$\lambda = \frac{N+\beta}{E(p_0\beta)} \bumpeq -\frac{N+\beta}{\log (p_0 \gamma' \beta)}, \quad \beta p_0 \bumpeq \exp\left[-\gamma - \frac{N+\beta}{\lambda} \right]. \tag{83}$$

If λ and x can be estimated from a sample, then β and p_0 can be determined by (82) and (83) and s can then be determined from (41), which gives

$$s + \beta = e^{-p_0\beta} / [p_0 E(p_0 \beta)] \bumpeq e^{-p_0\beta} / [-p_0 \log_e (p_0 \gamma' \beta)]. \tag{84}$$

In order to estimate λ and x from a sample, one could minimize χ^2, more or less as described above for H_3. For this purpose and for others it may be noted that, by (22),

$$V_N(n_r \mid H_6)/\mathscr{E}_N(n_r \mid H_6) \simeq 1 - \frac{r-1}{2r-1}\frac{1}{2^r}\binom{2r}{r} \quad (r \geqslant 2), \tag{85}$$

$$V_N(n_1 \mid H_6)/\mathscr{E}_N(n_1 \mid H_6) \simeq \frac{\lambda}{N+\beta} \tag{86}$$

$$\simeq \frac{2n_2'}{N} \quad \text{if} \quad x \simeq 1. \tag{87}$$

By comparing (85) with (65) we can get an idea of the smallness of the error arising when calculating χ^2 if (65) is used for hypotheses other than H_3.

Another method of estimating λ and x, rather less efficient, but easier, is the one analogous to that used by Fisher for H_3, namely, we may assume that the expected values of $N - n_1$ and of $S - n_1$ are equal to their observed values, i.e.

$$S - n_1 = \lambda \sum_{r=2}^{\infty} \frac{x^r}{r(r-1)} = \lambda[x + (1-x)\log_e(1-x)] = \lambda(1 - e^{-Y} - Y e^{-Y}), \tag{88}$$

$$N - n_1 = \lambda \sum_{r=2}^{\infty} \frac{x^r}{r-1} = -\lambda x \log_e(1-x) = \lambda Y(1 - e^{-Y}) = \lambda x Y, \tag{89}$$

$$(S - n_1)/(N - n_1) = Y^{-1} - (e^Y - 1)^{-1}, \tag{90}$$

where $x = 1 - e^{-Y}$. We may solve (90) iteratively, for Y, i.e. $Y = \lim_{n\to\infty} Y_n$, where $Y_1 = 0$ and, for $n = 1, 2, 3, \ldots$,

$$Y_{n+1}^{-1} = \frac{S - n_1}{N - n_1} + (e^{Y_n} - 1)^{-1}. \tag{91}$$

When λ and x are specified, the cumulative sums of $\mathscr{E}_N(n_r \mid H_6)$ can be found by making use of the approximation

$$\sum_t^{t \geqslant r} \frac{x^t}{t(t-1)} \simeq xE[-(r-1)\log_e x] - E(-r\log_e x) + \frac{x^r}{2r(r-1)}\left(1 + \tfrac{1}{6}\log_e x - \frac{1}{3r}\right), \tag{92}$$

which will be a very good approximation if the terms involving $\frac{1}{6}\log x$ and $\dfrac{1}{3r}$ are negligible (cf. equation (72)). If $(1-x)r$ is small while r is large, then we can prove the following approximation:

$$\sum_t^{t \geqslant r} \mathscr{E}_N(n_t \mid H_6) \simeq \lambda x\left\{\frac{1}{r} - (1-x)[1 - \gamma - \log_e(1-x) - \log_e r]\right\}. \tag{93}$$

If $1 - x$ is small but $(1-x)r$ is large, then

$$\sum_t^{t \geqslant r} \mathscr{E}_N(n_t \mid H_6) \simeq \frac{\lambda x^r}{(1-x)r^2}. \tag{93 A}$$

When in doubt about the accuracy of (93) and (93 A) it is best to use (92), the calculation of which is, however, ill-conditioned, so that the error integrals may be needed to several decimal places.

(vii) We now come to the 'less completely formulated' hypotheses. H_7 is discussed by Zipf, especially with $\zeta = 2$ and also in the slightly modified form

$$\mathscr{E}(n_r \mid H) \propto (r^2 - \tfrac{1}{4})^{-1}. \tag{94}$$

(See Zipf (1949, pp. 546–7), where there are further references, including ones to J. B. Estoup, M. Joos, G. Dewey and E. V. Condon.) Yule (1944, p. 55) refers to Zipf (1932) and objects to Zipf's word distributions on two grounds. First Yule asserts that the fits are unsatisfactory, and secondly he points out that (in our notation)

$$N = \mathscr{E}_N(\Sigma r n_r \mid H_7) = \infty \quad \text{if} \quad 1 < \zeta \leqslant 2,$$

while $$c_{2,0} = \int_0^1 p^2 f(p)\,dp = \mathscr{E}(p \mid H_7) = \mathscr{E}\left(\frac{\Sigma r(r-1)\,n_r}{N(N-1)} \middle| H_7\right) = \infty \quad \text{if} \quad 2 < \zeta \leqslant 3.$$

(viii) Yule's second objection to H_7 can be overcome by introducing a 'convergence factor', x^r, giving H_8. If H_7 is any good at all for any particular application then x will be fairly close to 1. It would be of interest to specify H_8 in terms of a density function, $f(p)$, by solving the simultaneous integral equations

$$\frac{\lambda x^r}{r^\zeta} = \frac{1}{r!}\int_0^\infty (Np)^r e^{-Np} f(p)\,dp \quad (r = 1, 2, 3, \ldots). \tag{95}$$

If $\zeta = 1$, then H_8 reduces of course to H_3.

(ix) H_9 is of interest mainly because it works so well in examples (ii) and (iii) of §8. Besides its formal similarity to H_8 with $\zeta = 2$, H_9 also resembles H_6, in virtue of equation (81). A disadvantage of not specifying $f(p)$ is that $V_N(n_r \mid H_9)$ cannot be conveniently worked out from (22), though it can always be estimated from (23) with considerably more work. Moreover, a correct specification of $f(p)$ is more fundamental than that of the expected values of the n_r's and is more likely to lead to a better understanding of the structure of the population.

In order to estimate λ and x from a sample, we could use either of the two methods discussed for H_3 and H_6, except that in the method of minimizing χ^2 it would perhaps be best to guess a formula for $V_N(n_r \mid H_9)$, after experimenting with formula (23). We shall not discuss this method further in this section. The second method consists in determining λ and x from the equations

$$N = \lambda \sum_{r=1}^{\infty} \frac{x^r}{r+1} = -\frac{\lambda}{x}[x + \log_e(1-x)], \tag{96}$$

$$S = \lambda \Sigma \frac{x^r}{r(r+1)} = \frac{\lambda}{x}[x + (1-x)\log_e(1-x)], \tag{97}$$

$$\frac{S}{N} = \frac{x + (1-x)\log_e(1-x)}{-x - \log_e(1-x)}. \tag{98}$$

x can be determined either by tabulating the right-hand side of (98) or by writing $x = 1 - e^{-Y}$ and determining Y from the equation

$$Y^{-1} = (1 - e^{-Y})^{-1} - (1 + S/N)^{-1}. \tag{99}$$

Y can be found iteratively by writing $Y = \lim_{n\to\infty} Y_n$, where $Y_1 = 1+N/S$, and, for $n = 1, 2, 3, \ldots$,

$$Y_{n+1}^{-1} = (1-e^{-Y_n})^{-1} - (1+S/N)^{-1}. \tag{100}$$

Then, by (96), we can find λ from $\lambda = \dfrac{xN}{Y-x}$. (101)

Having determined λ and x we may wish to test how well H_9 agrees with the sample. For this purpose we need to calculate cumulative sums of the expectations of the n_r's. This can be done by means of the approximation

$$\sum_{t}^{t\geqslant r} \frac{x^t}{t(t+1)} \simeq E(-r\log_e x) - \frac{1}{x}E[-(r+1)\log_e x] + \frac{x^r}{2r(r+1)}\left(1+\tfrac{1}{6}\log_e x - \frac{1}{3r}\right), \tag{102}$$

deducible from (92). If $(1-x)r$ is small while r is large, then we have the following approximation:

$$\sum_{t}^{t\geqslant r} \mathscr{E}(n_t \mid H_9) \simeq \lambda x\left\{\frac{1}{r} - (1-x)\left[1-\gamma-\log_e(1-x)-\log_e r\right]\right\}, \tag{103}$$

deducible from and of precisely the same form as (93). An idea of the closeness of this approximation can be obtained from example (ii) below. If $1-x$ is small but $(1-x)r$ is large, then

$$\sum_{t}^{t\geqslant r} \mathscr{E}(n_t \mid H_9) \simeq \frac{\lambda x^r}{(1-x)r^2}. \tag{103A}$$

When in doubt about the accuracy of (103) and (103A) it is best to use equation (102). (See the remarks following equation (93A).)

8. *Examples.* In each of the four examples given below we use at least two different methods of smoothing the data. One of these methods is, in each example, the graphical smoothing of $\sqrt{n_r}$ for the smaller values of r and another method is the fitting of one or other of the nine special hypotheses of §7. The discussion of these examples is by no means intended to be complete.

Example (i). *Captures of Macrolepidoptera in a light-trap at Rothamsted.* (Summarized from Williams's data (Corbet *et al.* 1943).) $N = 15{,}609$, $S = 240$.

r	n_r	n_r^{iv}	r	n_r	n_r^{iv}	r	n_r	n_r^{iv} (summed)†
1	35	40	11	2	3·5	21–30	18	15·5
2	11	20·0	12	2	3·2	31–50	16	18·0
3	15	13·2	13	5	3·0	51–70	17	11·4
4	14	9·9	14	2	2·8	71–100	8	11·2
5	10	7·9	15	4	2·6	101–150	9	11·8
6	11	6·6	16	3	2·4	151–200	7	7·4
7	5	5·6	17	3	2·3	201–500	12	16·1
8	6	4·8	18	3	2·1	501–1000	6	4·6
9	4	4·3	19	3	2·0	1001–∞	1	0·9
10	4	3·9	20	4	1·9	2349	1	—

† In future tables this word 'summed' will be taken for granted and omitted.

We now present the results of the calculations, followed by comments. (The columns headed n_r^{iv} in the table above are explained in these comments.)

r	n_r	n'_r	n''_r	n'''	n^{iv}_r	r^*	r^{**}	r^{***}	r^{****}
1	35	35	35	35	40	1·1	1·4	1·3	1
2	11	19·4	24·0	22·5	20·0	2·1	2·3	2·2	2
3	15	13·7	18·1	16·3	13·3	3·0	2·9	3·0	3
4	14	10·2	13·1	12·3	10·0	3·8	3·8	3·9	4
5	10	7·8	10·2	9·7	7·9	4·8	4·8	4·8	5
6	11	6·3	8·1	7·7	6·6	5·9	5·9	5·5	6
7	5	5·3	6·8	6·0	5·6	—	—	—	—

The function n'_r was obtained by plotting $\sqrt{n_r}$ against r for $1 \leqslant r \leqslant 20$ and smoothing for $1 \leqslant r \leqslant 7$ by eye, holding in mind the method of least squares. (See note (i) of §3.) n''_r was obtained in the same way, but an attempt was made to keep away from the graph of n'_r (except at $r = 1$) in order to find out how different a smoothing was reasonable. Next n'''_r was obtained by smoothing the cumulative sums $\sum_{t=1}^{r} tn_t$. Finally, n^{iv}_r is the function obtained by Fisher, i.e. using our hypothesis H_3 (equation (63)) with $\beta = 40\cdot 2$ and $x = 0\cdot 9974$. A more complete tabulation of n^{iv}_r is given in the first table. The 'summed' values of n^{iv}_r were calculated by means of equation (72). No statistical test is necessary to see that the fit of n^{iv}_r is very good. The values of r^* corresponding to the four smoothings of the data are denoted by r^*, r^{**}, r^{***} and r^{****} respectively. (Logically this gives r^* two different meanings.) ($r^{****} = 0\cdot 9974r$, by (2′) and (63).) In accordance with §3 we could force the r^*'s, etc., to be smooth. This has not been tried here. What is clear is that if H_3 is not accepted then most of the values of r^*, etc., are unreliable to within about 0·2 or 0·3. The approximate values of χ^2 given by (19) with $r = 7$ and assuming (65) are 10·9, 11·1, 9·4 and 11·7 respectively. The number of degrees of freedom is somewhere between 6 and 7. It seems safe to take it as 5 for n'''_r, 6 for n'_r and n''_r and 7 for n^{iv}_r. None of the values of χ^2 is particularly significant, though all are a bit large. The data can be blamed for the largeness of the values of χ^2, since n_2 is obviously much smaller than it ought to be. Of the four smoothings Fisher's seems to be the most likely to give the best approximations to the 'true expectations'. There is hardly anything to choose on the evidence of the sample, but Fisher's smoothing has the advantage of being analytically simple.

The most definite result of interest in this example does not depend much on the smoothing, namely, that the proportion of the population not represented by the species in the sample is about $(35 \pm 5)/15{,}609$. For the '± 5' see formula (65). Perhaps this standard error should be increased slightly, say from 5 to 8, to allow for the preference given to n^{iv}_r.

Formula (77), if it is applicable (i.e. if the truncated form, H_5, of H_3 is assumed), may be written $-\log_{10} p_0 = 1\cdot 18 + 0\cdot 011s$, so that if s were say 1000, then the smallest population frequency would be about 10^{-12}. This is mentioned only for its theoretical interest: it is an unjustifiable extrapolation to suppose that the distribution defined by H_5 would stand up to sample sizes large enough to demonstrate clearly the values of s and p_0. N would need to be of the order of $10/p_0$. The proposition which is made probable by the actual sample is that H_3 and H_5 (with the assigned values of the parameters) would give good fits to the values of n_r on other independent samples of 16,000 or less, i.e. that H_3 and H_5 provide good methods of smoothing the data. The cautious tone of this statement can be more fully justified by the following considerations.

If H_5 were reliable then it should be possible to use it to estimate the simpler measures of heterogeneity, such as $c_{2,0}$. Now we can see by (30) that $\hat{c}_{2,0} = 0{\cdot}03935$ and $\hat{c}_{3,0} \simeq 0{\cdot}0035$. (For the calculations, the complete data given by Williams must be used.) Hence, by (30A), it is reasonable to write $c_{2,0} = 0{\cdot}03935 \pm 0{\cdot}0007$. Let us then see what value for $c_{2,0}$ is implied by H_5. We have

$$\Sigma r(r-1)\, n_r^{\text{iv}} = \beta \Sigma (r-1)\, x^r = \beta x^2/(1-x)^2 = 0{\cdot}0243.$$

[As a check, $\displaystyle\int_0^\infty p^2 f(p)\, dp = \beta \int_{p_0}^\infty p\, e^{-\beta p}\, dp = \beta \int_{p_0\beta}^\infty q\, e^{-q} dq \simeq \beta^{-1} = 0{\cdot}025.$]

Clearly then H_5 cannot be used to estimate $c_{2,0}$. It would be true if misleading to say that H_5 is decisively disproved by the data. Similar remarks would apply in the examples below.

Example (ii). *Eldridge's statistics for fully inflected words in American newspaper English.* Eldridge's statistics (1911) are summarized by Zipf (1949, pp. 64 and 25). We give a summary of Zipf's summary in column (ii) below; more fully in the second table. $N = 43{,}989$, $S = 6{,}001$.

In this example the values of n_r for $r \leqslant 10$ are much larger than in example (i), so we have far more confidence in the smoothing that is independent of particular hypotheses. We shall present some of the numerical calculations in columns and then make comments on each column. We may assert at once, however, by equations (7), (8) and (9), that the proportion of the population represented by the sample is close to $1 - n_1/N = 14/15$. If a foreigner were to learn all 6001 words which occurred in the sample he would afterwards meet a new word at about 6·7 % of words read. If he learnt only $S - n_1 = 3025$ words he would meet a new word about 11·6 % of the time. The corresponding results for word-roots rather than for fully inflected words would be of more interest to a linguist.

(i) r	(ii) n_r	(iii) $\sqrt{n_r}$	(iv) b_r say	(v) $-\Delta$	(vi) rb_r	(vii) rb'_r say	(viii) b'_r	(ix) b'^2_r	(x) n'_r	(xi) r^*
1	2976	54·5	54·5	21·8	54·5	54·5	54·5	2976	2961	0·73
2	1079	32·7	32·7	10·0	65·4	65·4	32·7	1079	1075	1·4
3	516	22·7	22·7	5·7	68·1	67·8	22·6	511	509	2·4
4	294	17·1	17·0	2·8	68·0	70·2	17·5	206	305	3·4
5	212	14·6	14·2	1·8	71·0	72·6	14·5	210	209	4·4
6	151	12·3	12·4	1·5	74·4	74·7	12·4	154	153	5·4
7	105	10·2	10·9	1·3	76·3	76·3	10·9	119	118	6·2
8	84	9·2	9·6	1·2	76·8	76·8	9·6	92	91	—
9	86	9·3	8·4	1·1	75·6	75·6	8·4	71	70	—
10	45	6·7	7·3	—	—	—	—	—	—	—

(i) and (ii). We first consider the values of r only as far as $r = 10$. For larger values of r the smoothing could be done by using k-point smoothing formulae with $k \simeq 2\sqrt{r}$.

(iii) Each entry in this column has standard error of about $\frac{1}{2}$, so one place of decimals is appropriate.

(iv) This column was obtained by smoothing a graph of column (iii) by eye. Experiments with the five-point smoothing formula did not give quite as convincing results. For the five-point smoothing formula, see, for example, Whittaker & Robinson (1944, § 146). For the present application it would be $\sqrt{n'_r} = \sqrt{n_r} - \frac{3}{35}\Delta^4(\sqrt{n_r})$ $(r = 3, 4, 5, \ldots)$.

(v) This column of differences is given as a verification of the smoothness of column (iv). In fact minor adjustments were made in column (iv) in order to improve the smoothness of column (v).

(vi) The numbers b_r of column (iv) are roughly proportional to r^{-1}. This fact suggests that rb_r should be formed and smoothed again in order to improve the smoothing of $\sqrt{n_r}$ still further. This process is of course distinct from assuming that rb'_r should be constant, where the function b'_r is a smoothing of the function b_r.

(vii) and (viii) These columns have already been partly explained. The purpose of this improvement in the smoothing is more for the sake of the ratios n'_{r+1}/n'_r than of the n'_r themselves.

(ix) Where the smoothing of $\sqrt{n_r}$ had no noticeable effect we have taken $b'^2_r = n_r$. It is clearly typical that $b'^2_1 = n_1$, since the eye-smoothing is unlikely to affect n_1 convincingly. Therefore if the smoothing is tested by means of a chi-squared test it will be reasonable to subtract about two degrees of freedom.

(x) We have scaled up column (ix) so as to force $\sum_{r=1}^{9} rn'_r = \sum_{r=1}^{9} rn_r$. We can then assume $N' = N$, convenient for applications of §6. Note that $\sum_{r=1}^{9} k_r(n'_r - n_r)^2/n'_r = 6{\cdot}5$, so that χ^2, given by (19) and accepting (65) as a good enough approximation, is not significant on eight degrees of freedom. Thus our smoothing is satisfactory, though there may be other satisfactory smoothings.

(xi) r^* is obtained from formula (2'). The larger is r the larger is the standard error of r^*. We may get some idea of the error by means of an alternative smoothing. The standard error of 1^* can be very roughly calculated by an *ad hoc* argument, inapplicable to say 5^*. We may reasonably say that the variance of $2n'_2/n'_1$ with respect to all eye-smoothings will be about the same as that obtained by regarding n'_2 and n'_1 as independent random variables with variances circumscribed by the inequalities (26) and (27), or nearly enough, defined by (65). Now if w and z are independent random variables with expectations W and Z, we have

$$\delta\left(\frac{w}{z}\right) = \frac{\delta w}{Z} - \frac{W\,\delta z}{Z^2},$$

and hence, to a crude approximation,

$$V\left(\frac{w}{z}\right) \simeq \frac{V(w)}{Z^2} + \frac{W^2 V(z)}{Z^4},$$

i.e.

$$\frac{V(w/z)}{W^2/Z^2} \simeq \frac{V(w)}{W^2} + \frac{V(z)}{Z^2}. \tag{104}$$

It follows that

$$\frac{V(1^*)}{1^{*2}} = \frac{V(2n'_2/n'_1)}{(2n'_2/n'_1)^2} \simeq \frac{1}{k_1 n'_1} + \frac{1}{k_2 n'_2}, \tag{105}$$

so that $V(1^*) = 0{\cdot}73^2 \times 0{\cdot}0010 = 0{\cdot}00052$ and $1^* = 0{\cdot}73 \pm 0{\cdot}023$.

(xii) (see the second table). An analytic smoothing which is remarkably good for $r \leqslant 15$ is given by $n''_r = S/(r^2+r)$. For larger values of r there is a serious discrepancy, since $\sum_{r=16}^{\infty} n''_r = 374$ while $\sum_{r=16}^{\infty} n_r = 297$. It is clear without reference to the sample that n''_r cannot be satisfactory for sufficiently large values of r, since $\Sigma rn''_r = \infty$ instead of being equal to N.

(i) r	(ii) n_r	(x) n'_r	(xii) n''_r	(xiii) n'''_r	(xi) r^*	(xiv) r^{**}
1	2976	2961	3000	3008	0·73	0·67
2	1079	1075	1000	1002	1·4	1·5
3	516	509	500	500	2·4	2·4
4	294	305	300	301	3·4	3·3
5	212	209	200	201	4·4	4·3
6	151	153	143	144	5·4	5·2
7	105	118	107	108	6·2	6·2
8	84	91	83	84	—	7·2
9	86	70	67	67	—	8·2
10	45	—	55	55	—	—
11–15	156	—	170	170	—	—
16–20	76	—	89	89	—	—
21–30	78	—	92	92	—	—
31–40	34	—	47	47	—	—
41–50	28	—	29	28	—	—
51–60	10	—	19	19	—	—
61–∞	71	—	98	90	—	—
4290	1	—	—	—	—	—

(xiii) The fit can be improved by writing $n'''_r = \lambda x^r/(r^2+r)$ as in equation (55), i.e. using hypothesis H_9. We find by equations (100) and (101) that $\lambda = 6017{\cdot}4$ and $x = 0{\cdot}999667$. Column (xiii) can then be easily calculated directly for $r \leqslant 10$ and by use of (102) or (103) for $r > 10$. ((103) gives the correct values for n'''_9 and n''''_{10}, to the nearest integer, and it gives $\sum\limits_{61}^{\infty} n''''_r = 89{\cdot}96$, as compared with 89·90 when (102) is used.) Note that $\sum\limits_{r=16}^{\infty} n'''_r = 365$, which implies an improvement on n''_r but is still significantly too large. A better fit could be obtained by the method of minimum χ^2 or by using some simple convergence factor other than x^r, such as e^{-ar-br^2} with $a > 0$, $b > 0$.

(xiv) r^{**} is defined as $(r+1)\,n_{r+1}/n_r$ and is equal to $r(r+1)/(r+2)$. This column may be compared with column (xi). The agreement looks fairly good. It is by no means clear which of the two columns gives more reliable estimates of the 'true' values of r^* for $r \leqslant 7$. Column (x) is a better fit to Eldridge's data for $r \leqslant 9$ (and could be extended to be a better fit for all r) than is column (xiii) but is not as smooth. Columns (xii) and (xiii) would be preferable if some theoretical explanation of the analytic forms could be provided. Such an explanation might also show why the fit is not good for large r, even with the convergence factor x. The limitation on r in equation (46) may be relevant.

If H_9 is true, the population parameter $c_{2,0}$, given by (31), can be expressed in the form

$$\frac{\lambda}{N^{(2)}}\sum_{r=1}^{\infty}\frac{r-1}{r+1}x^r = \frac{\lambda}{N^{(2)}}\left[\frac{x}{1-x} - 2 + \frac{2}{x}\log_e(1-x)\right]. \tag{106}$$

Formula (106) would give $c_{2,0} = 0{\cdot}00928$, but this value is probably a bad over-estimate since n'''_r is too large for large r and the terms of $\dfrac{\lambda}{N^{(2)}}\Sigma\dfrac{r-1}{r+1}x^r$ for large r make most of the contribution. Similarly, $\hat{c}_{2,0}$, given by (30), depends mainly on the larger values of r represented in the sample, but Zipf's summary of Eldridge's data is not complete enough to calculate $\hat{c}_{2,0}$. Similarly, assuming H_9, the entropy, $c_{1,1}$, could be estimated from equation

(39), and this method could be expected to give close agreement with the correct value, since $c_{1,1}$ does not depend so very much on the more frequent species. But I have not obtained a closed formula, resembling (106) for example, and the arithmetic required if no closed formula is available would be heavy. The estimation of measures of heterogeneity will be discussed again under example (iii).

Example (iii). *Sample of nouns in Macaulay's essay on Bacon.* (Taken from Yule (1944) Table 4·4, p. 63.) $N = 8045$, $S = 2048$.

r	n_r	r	n_r	r	n_r	r	n_r	r	n_r
1	990	11	24	21	1	31	2	41	1
2	367	12	19	22	4	32	1	45	2
3	173	13	10	23	7	33	1	48	1
4	112	14	10	24	2	34	1	57	1
5	72	15	13	25	1	35	1	58	1
6	47	16	3	26	5	36	1	65	1
7	41	17	10	27	3	37	1	76	1
8	31	18	7	28	4	38	2	81	1
9	34	19	6	29	1	39	4	89	1
10	17	20	5	30	3	40	1	255	1

As in example (ii) we can state some conclusions at once, without doing the smoothing. If our foreigner learns all 2048 nouns that occur in the sample his vocabulary will represent all but (12·3 ± 0·5) % of the population, assuming formulae (9) and (65) or (87). If he learns only 1058 nouns his vocabulary will still represent all but $(n_1+2n_2)/N = 19{\cdot}3$ % of the population.

We now present three different smoothings corresponding precisely to those of example (ii).

r	n_r	n'_r	n''_r	n'''_r	r^*	r^{**}	r^{***}	$\frac{d}{dr}\log_{10} n'_r$	$\frac{d}{dr}\log_{10} n'''_r$	$g_r \log_{10} e$
1	990	990	1024	1060	0·74	0·67	0·66	−0·50	−0·65	0·184
2	367	367	341	350	1·4	1·5	1·5	−0·30	−0·37	0·401
3	173	173	170	174	2·6	2·4	2·4	−0·24	−0·26	0·545
4	112	112	102	103	3·4	3·3	3·3	−0·17	−0·20	0·654
5	72	76	68	68	4·4	4·3	4·3	−0·15	−0·16	0·741
6	47	56	49	48	5·3	5·2	5·1	−0·12	−0·14	0·813
7	41	42	35·5	36	6·5	6·2	6·1	−0·11	−0·12	0·876
8	31	34	28·5	28	7·3	7·2	7·1	−0·10	−0·11	0·930
9	34	27	22·7	22	8·2	8·2	8·1	−0·09	−0·10	0·978
10	17	22	18·4	18	—	9·2	9·1	−0·08	−0·09	1·021
11	24	18·5	15·5	15	—	10·2	10·1	—	—	—
12	19	16·0	13·1	12	—	11·1	11·0	—	—	—
13	10	13·7	11·3	10	—	12·1	12·0	—	—	—
14	10	10·9	9·7	9	—	13·1	13·0	—	—	—
15	13	9·6	8·5	8	—	14·1	14·0	—	—	—
16–20	31	32·5	30·5	27	—	—	—	—	—	—
21–30	31	—	31·5	26	—	—	—	—	—	—
31–50	19	—	25·9	19	—	—	—	—	—	—
51–100	6	—	19·9	11	—	—	—	—	—	—
101–∞	1	—	20·3	3·6	—	—	—	—	—	—
255	1	—	—	—	—	254	252	—	—	—

$\sqrt{n'_r}$ was obtained by smoothing $\sqrt{n_r}$ graphically.

$n_r'' = S/(r^2+r)$. It is curious that this should again give such a good fit for values of r that are not too large ($r \leqslant 30$). The sample is of nouns only and, moreover, Yule took different inflexions of the same word as the same.

$n_r'''' = \lambda x^r/(r^2+r)$, where $\lambda = 2138{\cdot}90$, $x = 0{\cdot}991074$, the values being obtained from (100) and (101) as in example (ii).

The expressions $\sum_{r=1}^{15} (n_r' - n_r)^2/n_r'$, etc., take the values 9·5, 21·2 and 27·3. The values of χ^2 would be about 2 or 3 larger. (See (19), (26), (27), (65).) There is no question of accepting n_r'' for $r > 50$ but it is better than n_r'''' for $r \leqslant 15$. When $r \leqslant 9$ the values of r^* and r^{**} (and therefore of r^{***}) show good agreement except for $r = 1$ and $r = 7$. If the analytic smoothings had not been found, the value of 6^* would have been smoothed off, with repercussions on the function n_r'. The discrepancy in 1^* must be attributed either to a fault in the value of n_1'''' (and therefore in H_9) or must be blamed on n_1 (i.e. on sample variation). If I had not noticed the analytic smoothings I would have asserted that $1^* = 0{\cdot}74$ with a standard error of something like 0·04. (See equation (105).)

We now consider two of the measures of heterogeneity in the population, namely, $c_{2,0}$ and $c_{1,1}$. By (30) we can see that $\hat{c}_{2,0} = 0{\cdot}00272$, agreeing with Yule (1944, p. 57). Also $\hat{c}_{3,0} = 0{\cdot}00003957$, so that by (30A) we may reasonably write $c_{2,0} = 0{\cdot}00272 \pm 0{\cdot}00013$. Assuming H_9 to be valid for $r \leqslant 30$, we may also estimate $c_{2,0}$ by $\tilde{c}_{2,0}\,(30)$ as in equation (33). We have, in a self-explanatory notation,

$$\tilde{c}_{2,0}(30 \mid H_9) = \frac{1}{N^{(2)}}\left\{\lambda \sum_{1}^{30} \frac{r-1}{r+1}\, x^r + \sum_{30}^{\infty} r^{(2)} n_r\right\}. \tag{107}$$

Now, as in (72),

$$\sum_{31}^{\infty} \frac{r-1}{r+1}\, x^r \simeq \frac{x^{31}}{1-x} - \frac{2}{x}\left\{E(-32\log_e x) + \frac{x^{32}}{64}\,(1 + \tfrac{1}{6}\log_e x - \tfrac{1}{192})\right\} = 82{\cdot}924.$$

But, as in (106), $\sum_{1}^{\infty} \frac{r-1}{r+1} x^r = 99{\cdot}501$, so that $\sum_{1}^{30} \frac{r-1}{r+1} x^r = 16{\cdot}577$. It follows from (107) that $\tilde{c}_{2,0}(30 \mid H_9) = 0{\cdot}00246$. This is about two standard errors below its expected value, based on the simple unbiased statistic $\hat{c}_{2,0}$. The discrepancy may again be attributed to the large value of n_1''''. If, instead of n_r'''', the smoothing n_r'' is accepted for $r \leqslant 30$, we would get $\tilde{c}_{2,0}(30) = 0{\cdot}00267$. (It was in order to obtain this comparison that we calculated $\tilde{c}_{2,0}(30 \mid H_9)$ rather than $\tilde{c}_{2,0}(50 \mid H_9)$. The fit of n_r'' deteriorates at about $r = 30$.)

The last three columns of the table are related to the estimation of the entropy, $c_{1,1}$. (See equation (40) and the remarks following it.) $\frac{d}{dr}\log_{10} n_r'$ was obtained graphically for $r = 1$, 2 and 3 by numerical differentiation for $r = 3, 4, \ldots, 10$. (The graphical and numerical values agreed to two decimal places for $r = 3$.) The column $\frac{d}{dr}\log_{10} n_r''''$ was of course calculated as $\log_{10} x - \left(\frac{1}{r} + \frac{1}{r+1}\right)\log_{10} e$. The crude estimate of the 'entropy to base 10' or 'entropy expressed in decimal digits' is $\log_{10} N - \frac{1}{N}\sum_r r n_r \log_{10} r = 2{\cdot}968$ decimal digits. If n_r' is accepted for $r = 1, 2, 3, \ldots, 10$ we find that

$$\tilde{c}_{1,1}(10) = \log_{10} N - \frac{1}{N}\left\{\sum_{r=1}^{10} r n_r'\left(g_r \log_{10} e + \frac{d}{dr}\log_{10} n_r'\right) + \sum_{r=11}^{\infty} r n_r \log_{10} r\right\} = 3{\cdot}051 \text{ decimal digits.}$$

We shall next calculate $\tilde{c}_{1,1}(50 \mid H_9)$, using another self-explanatory notation. Since, by Jeffreys & Jeffreys (1946, §15·05),

$$g_r \sim \log_e r + \frac{1}{2r} - \frac{1}{12r^2} - \dots,$$

it can be seen that

$$\tilde{c}_{1,1}(50 \mid H_9) \simeq \log_{10} N - \frac{1}{N}\left\{\sum_{r=1}^{10} rn_r''' \left(g_r \log_{10} e + \frac{d}{dr}\log_{10} n_r'''\right)\right.$$
$$\left. + \sum_{11}^{50} rn_r''' \log_{10} r + \log_{10} x \sum_{11}^{50} rn_r''' - \frac{3\log_{10} e}{2}\sum_{11}^{50} n_r''' + \sum_{51}^{\infty} rn_r \log_{10} r\right\}$$
$$= 3{\cdot}192 \text{ decimal digits,}$$

as we may see by means of rather heavy calculations, using the last column of the table, together with equations (72), (74) and (92). The crude estimate of $c_{1,1}$ is the smallest of the three. This is not surprising, since the crude estimate is always too small in the special case of sampling from a population of s species all of which are equally probable.

Example (iv). *Chess openings in games published in the British Chess Magazine*, 1951. For the purposes of this example we arbitrarily regard the openings of two games as equivalent only if the first six moves (three white and three black) are the same and in the same order in both games. $N = 385$, $S = 174$.

r	n_r	n_r'	n_r''	n_r'''	r^{**}
1	126	126	126	126	0·39
2	22	22	24·6	24	1·0
3	5	7·6	8·1	8	2·0
4	4	4·8	4·0	4	3·0
5	3	3·2	2·4	2·4	4·0
6	4	2·6	1·6	1·6	5·0
7	0	2·2	1·1	1·14	6·0
8	3	—	0·85	0·86	7·0
9	1	—	0·66	0·67	8·0
10	1	—	0·52	0·53	9·0
11	0		3·97	4·80	
13	1				
14	1				
16	1				
23	1				
36	1				
∞	—				

$\sqrt{n_r'}$ was obtained by graphical smoothing of $\sqrt{n_r}$.

n_r'' was obtained by assuming H_6 (see equation (52)), i.e. $n_r'' = \mathscr{E}_{385}(n_r \mid H_6)$, where the parameters x and λ were obtained from (91) and (89). These gave $x = 0{\cdot}99473$, $\lambda = 49{\cdot}635$ and n_r'' for $r \geqslant 2$ is then given by (81). Next p_0 was determined as $0{\cdot}00011304 = 1/8846$ by using equation (80). Then (82) gave $\beta = 2{\cdot}040$, so that, in accordance with (52) and (74),

$$f(p) \simeq \begin{cases} 0{\cdot}128p^2 e^{-2{\cdot}040p} & (p > 1/8846), \\ 0 & (p < 1/8846). \end{cases}$$

Finally, equation (84) gives $s = 1132$. This then is the estimate of the total number of openings in the population, though the sample is too small to put any reliance in it.

$$n_r'''(r \geqslant 2) \text{ is simply } (S - n_1)/(r^2 - r) = 48/(r^2 - r).$$

This is just as good a fit as n_r''. It gives an infinite value to $c_{2,0}$, but this is not as serious an objection as it sounds since H_6 would also give quite the wrong value for $c_{2,0}$. (Cf. the concluding remarks in the discussion of example (i).)

We list in the table the values of r^* corresponding to n_r'', calling the values r^{**} in conformity with the convention of the present section. Clearly $r^{**} = (r-1)x$ when $r \geqslant 2$. Thus the average population frequency of the 126 openings that each occurred once only in the sample is 0·39/385 = 0·001.

A player who learnt all 174 openings would expect to recognize about 67 % of future openings for the same population, assuming that the sample was random. If he learnt the 48 openings that each occurred twice or more in the sample the percentage would drop to 55 % and if he learnt the 26 that occurred three times or more the percentage would drop to 49 %. (See formula (6′).)

9. *Index of notations having a fixed meaning.*

§1. N, n_r (but see also §2), n_0, q_r, r^* (as a definition of the asterisk, but there is a slight change of convention in §8), n_r' (here again there is a slight change in §8), $\mathscr{E}(\ )$, $V(\ )$.

§2. $s, p_\mu, H(p_1, p_2, \ldots, p_s) = H, \mathscr{E}_N, \mu'_{r,t,N}$.

§3. N'.

§5. $x_{\mu,r} = x_\mu, a_\mu$.

§6. $c_{m,n}, \hat{c}_{m,0}, \tilde{c}_{m,0}, \tilde{c}_{m,0}(t), \gamma, g_r, \hat{\tilde{c}}_{1,1}, \tilde{c}_{m,1}(t)$.

§7. $p, f(p), p_0, H_1$ to H_9, $E(\ ), k_r, S, \gamma'$.

REFERENCES

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika*, **35**, 246–54.

Anscombe, F. J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37**, 358–82.

Bartlett, M. S. (1936). The square root transformation in the analysis of variance. *J. R. Statist. Soc. Suppl.* **3**, 68–78.

Chambers, E. G. & Yule, G. U. (1942). Theory and observation in the investigation of accident causation. (Including discussion by J. O. Irwin and M. Greenwood.) *J. R. Statist. Soc. Suppl.* **7**, 89–109.

Corbet, A. S., Fisher, R. A. & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58.

Eldridge, R. C. (1911). *Six Thousand Common English Words*. Buffalo: The Clements Press. (Mentioned in Zipf (1949).)

Fletcher, A., Miller, J. C. P & Rosenhead, L. (1946). *An Index of Mathematical Tables*. London: Scientific Computing Service.

Good, I. J. (1950*a*). A proof of Liapounoff's inequality. *Proc. Camb. Phil. Soc.* **46**, 353.

Good, I. J. (1950*b*). *Probability and the Weighing of Evidence*. London: Charles Griffin.

Goodman, L. A. (1949). On the estimation of the number of classes in a population. *Ann. Math. Statist.* **20**, 572–9.

Greenwood, M. & Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. R. Statist. Soc.* **83**, 255–79.

HARDY, G. H. (1949). *Divergent Series*. Oxford: Clarendon Press.

JAHNKE, E. & EMDE, F. (1933). *Funktionentafeln mit Formeln und Kurven*, 2nd ed. Leipzig and Berlin.

JEFFREYS, H. (1948). *Theory of Probability*, 2nd ed. Oxford: Clarendon Press.

JEFFREYS, H. & JEFFREYS, B. S. (1946). *Methods of Mathematical Physics*. Cambridge University Press.

JOHNSON, W. E. (1932). Appendix (edited by R. B. Braithwaite) to 'Probability: deductive and inductive problems'. *Mind*, **41**, 421–3.

NEWBOLD, E. M. (1927). Practical applications of the statistics of repeated events, particularly to industrial accidents. (Including discussion by M. Greenwood, D. R. Wilson, M. Culpin, E. Farmer and L. Isserlis.) *J. R. Statist. Soc.* **90**, 487–547 (esp. Appendix, pp. 518–35).

PRESTON, F. W. (1948). The commonness, and rarity, of species. *Ecology*, **29**, 254–83.

SHANNON, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423.

SIMPSON, E. H. (1949). Measurement of diversity. *Nature, Lond.*, **163**, 688.

USPENSKY, J. V. (1937). *Introduction to Mathematical Probability*. New York: McGraw Hill.

WHITTAKER, E. T. & ROBINSON, G. (1944). *The Calculus of Observations*, 4th ed. London and Glasgow: Blackie.

WHITTAKER, E. T. & WATSON, G. N. (1935). *A course of Modern Analysis*, 4th ed. Cambridge University Press.

YULE, G. U. (1944). *Statistical Study of Literary Vocabulary*. Cambridge University Press.

ZIPF, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press. (Mentioned in Yule (1944) and Zipf (1949.)

ZIPF, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley Press.