THE NUMBER OF NEW SPECIES, AND THE INCREASE IN POPULATION COVERAGE, WHEN A SAMPLE IS INCREASED

By I. J. GOOD and G. H. TOULMIN

A sample of size N is drawn at random from a population of animals of various species. Methods are given for estimating, knowing only the contents of this sample, the number of species which will be represented r times in a second sample of size λN; these also enable us to estimate the number of different species and the proportion of the whole population represented in the second sample. A formula is found for the variance of the estimate; when $\lambda > 2$, this variance becomes in general very large, so that the estimate is useless without some modification. This difficulty can be partly overcome, at least for $\lambda < 5$, by using Euler's method with a suitable parameter or the methods described by Shanks (1955) to hasten the convergence of the series by which the estimate is expressed. The methods are applied to samples of words from *Our Mutual Friend*, to an entomological sample, and to a sample of nouns from Macaulay's essay on Bacon.

1. Introduction

We present here a further development of the theory expounded by Good (1953); that paper will be referred to, for brevity, by the letter G throughout.

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

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

We write

$$d = \sum_{r=1}^{\infty} n_r,$$

the total number of distinct species in the sample. It is convenient (though, as was pointed out in G, not essential) to suppose that the total number of distinct species in the population is a known finite number s, so that we can calculate

$$n_0 = s - d, \tag{2}$$

the number of species not represented in the sample. If the actual value of s is not known, all our results will remain true if it is arbitrarily assumed to be any sufficiently large number. As in G (p. 237), the larger n_1 is, the more applicable our results are. In G it was shown that certain properties of the population could be deduced approximately from the sample frequencies n_r; in particular, the total *coverage* of the sample (i.e. the proportion of the population represented in the sample, which is the sum of the population frequencies p_μ of the species represented) is approximately

$$1 - \frac{\mathscr{E}(n_1)}{N} \simeq 1 - \frac{n_1}{N}, \tag{3}$$

rovided n_1 is large (G, formula (9)).†

† $\mathscr{E}(n_1)$ is the expected value of the random variable n_1 when our basic sample of N specimens is taken at random. We shall use the same symbol n_1 both for this random variable and for a particular value of it.

We now contemplate taking a *second sample*, of size λN. We describe this as the 'second sample', even though it may be (and in practice probably will be) an enlargement of the basic sample; in this case, of course, $\lambda \geqslant 1$. If the second sample is not an enlargement of the basic sample, it will be termed *independent*; this word may be interpreted in its probabilistic sense, provided that the true statistical hypothesis specifying the population frequencies is momentarily regarded as 'given'. Except in §4 our results apply to both enlargements and independent second samples.

We may now wish, for example:

(*a*) To find the expected coverage of the second sample.

(*b*) To find the expected number of distinct species in the second sample.

(*c*) To find (roughly) the variances of estimates of population parameters which might be made from the second sample.

(*d*) To estimate the term $\mathscr{E}_{2N}(n_{2r} \mid H)$ in formula (22) of G for the variance of n_r.

Results of this type may enable us to decide whether it is worth enlarging our sample, and to what extent, depending on the purposes for which it is required.

For example, consider a teacher of languages who wishes to base his teaching on the population frequencies of words. He will wish to estimate what size of vocabulary should be learnt by a student in order to decrease the need for reference to a dictionary below a certain frequency. It was shown in G how a sample can be used to make such an estimate. The present paper shows in what way the sample can also be used in order to help the decision of whether to carry out more sampling. For instance, in example (iii) of §6 below, the 2048 words of the basic sample had an expected coverage of 87·3 %, and we find that if the sample size were doubled, then the same expected coverage could be obtained by selecting only 1780 words. More work by the teacher means less for the student.

Similarly, an entomologist will often want to know whether to increase a basic sample, and will be able to base his decision largely on the expected number of new species that will be provided by a given amount of sampling. Example (ii) of §6 is an instance of such an application.

Let $n_r(\lambda)$ be the random variable whose value is the number of distinct species represented exactly r times in the second sample.†

We first consider a method that may appeal to statisticians who are accustomed to fit distributions by the method of moments. The method will not, however, be used in our examples if only because of the enormous amount of calculation that it requires. We begin by stating a lemma that is presumably well known, although we cannot give a reference.

LEMMA. (Determination of a set of numbers whose 'factorial moments' are specified.) If

$$b_i = \sum_{r=0}^{\infty} r^{(i)} a_r \quad (i = 0, 1, 2, \ldots), \tag{4}$$

then

$$a_r = \frac{1}{r!} \sum_{i=0}^{\infty} \frac{(-1)^i b_{r+i}}{i!}, \tag{5}$$

at any rate if $a_r = 0$ for all sufficiently large r.‡ Problems (*a*), (*b*) and (*d*) above reduce to the estimation of the numbers $\mathscr{E}(n_r(\lambda))$ for certain values of r when values of $n_r(1) = n_r$ are

† It is convenient here to depart slightly from the notation of G. The number which we write as $\mathscr{E}(n_r(\lambda))$ would there have been denoted by $\mathscr{E}_{\lambda N}(n_r)$. Note that in the case of an enlarged sample, $n_r(\lambda)$ is to be considered as varying as the whole enlarged sample is varied at random, not merely the $(\lambda - 1)N$ additional specimens, so that this correspondence of notation still holds.

‡ For a proof under more general conditions, see the Appendix p. 62 below.

observed; the results are given in equations (22), (23) and (32), respectively. The same is true of problem (*c*) when the variances concerned can be expressed in terms of the $n_r(\lambda)$; thus equations (30A) and (31) of G show that this is the case for Yule's 'characteristic', which is an estimate of $\sum_{\mu=1}^{s} p_\mu^2$.

Now
$$\frac{1}{N^{(i)}} \sum_{r=0}^{\infty} r^{(i)} n_r$$
is an unbiased estimate of $c_i = \sum_{\mu=1}^{s} p_\mu^i$ (see, for example, G, p. 245). Similarly

$$\mathscr{E}\left\{\frac{1}{(\lambda N)^{(i)}} \sum_{r=0}^{\infty} r^{(i)} n_r(\lambda)\right\} = \sum_{\mu=1}^{s} p_\mu^i. \tag{6}$$

It may therefore seem reasonable to assume

$$\sum_{r=0}^{\infty} r^{(i)} \mathscr{E}(n_r(\lambda)) \simeq \frac{(\lambda N)^{(i)}}{N^{(i)}} \sum_{r=0}^{\infty} r^{(i)} n_r \left[\simeq \lambda^i \sum_{r=0}^{\infty} r^{(i)} n_r\right], \tag{7}$$

and then to solve for $\mathscr{E}(n_r(\lambda))$ by using the lemma.

In spite of the theoretical interest of this method it seems likely that it is not really adequate. For it depends too much on the estimates of the higher population moments, c_i, and these estimates are subject to large sampling errors. We have therefore not investigated any numerical examples. Instead of proceeding via the factorial moments, we find directly the following relation between $\mathscr{E}(n_r(\lambda))$ and the $\mathscr{E}(n_r)$:

$$\mathscr{E}(n_r(\lambda)) \simeq \lambda^r \sum_{i=0}^{\infty} (-1)^i \binom{r+i}{r} (\lambda - 1)^i \mathscr{E}(n_{r+i}) \tag{8}$$

for any integer $r \geqslant 0$ (§2, equation (16)). If we assume that

$$\mathscr{E}(n_{r+i}) \simeq n_{r+i} \tag{9}$$

or
$$\mathscr{E}(n_{r+i}) \simeq n'_{r+i}, \tag{10}$$

where the numbers $n'_1, n'_2, n'_3, \ldots$ are obtained by smoothing the numbers $n_1, n_2, n_3, \ldots$ (see G, §§3, 7, 8), we can estimate the values of $\mathscr{E}(n_r(\lambda))$.

We point out here that the series (8) is not really infinite: for

$$\mathscr{E}(n_{r+i}) = 0 \quad \text{whenever} \quad r+i > N, \tag{11}$$

and the upper limit of summation could therefore be replaced by $N-r$. If $\lambda > 2$, a practical difficulty arises, in that the factor $(\lambda-1)^i$ increases rapidly with i, and so attaches great weight to terms for which $\mathscr{E}(n_{r+i})$ is small and therefore is liable to a large percentage error when estimated from the basic sample. It seems to be practicable to overcome this difficulty, at least for moderate values of λ (say $\lambda \leqslant 5$), by using a summation method to make the series (8) converge rapidly; it is shown in §5 that Euler's method with a suitably chosen parameter q is convenient. We have not, however, been able to justify this procedure by finding a useful error term for the partial sums of the new series obtained.

We mention here two possibilities which we have not investigated practically. First, it may be possible to reach larger values of λ in two (or more) stages: e.g. to estimate $\mathscr{E}(n_1(4))$ we might, instead of using (8) directly with $\lambda = 4$, first estimate $\mathscr{E}(n_1(2))$, $\mathscr{E}(n_2(2))$,

$\mathscr{E}(n_3(2)), \ldots$, then smooth the values obtained, and again apply (8) with $\lambda = 2$ to estimate $\mathscr{E}(n_1(4))$. Secondly, (8) with $1/\lambda$ in place of λ and n_r, $n_r(\lambda)$ interchanged might be used as a check on the results obtained, and this might provide a new method of smoothing.

We are much indebted to Dr J. Wishart for several suggestions and corrections.

2. Estimation of $\mathscr{E}(n_r(\lambda))$

Let $p_\mu (\mu = 1, 2, \ldots, s)$ be the population frequencies of the s species. As in G, equation (10),

$$\mathscr{E}(n_r) = \sum_{\mu=1}^{s} \binom{N}{r} p_\mu^r (1-p_\mu)^{N-r}. \tag{12}$$

(In G, the left-hand side is written as $\mathscr{E}_N(n_r | H)$. As explained above, we use the symbol n_r only with reference to the basic sample, of size N; we omit the H, which refers to the hypothesis that the population frequencies are $\{p_\mu\}$, because we shall not be concerned with expectations on any other hypothesis.) For the second sample, we have similarly, assuming $p_\mu < \frac{1}{2}$ for all μ,

$$\begin{aligned}
\mathscr{E}(n_r(\lambda)) &= \sum_{\mu=1}^{s} \binom{\lambda N}{r} p_\mu^r (1-p_\mu)^{\lambda N - r} \\
&= \sum_{\mu=1}^{s} \binom{\lambda N}{r} p_\mu^r (1-p_\mu)^{N-r} \left(1 + \frac{p_\mu}{1-p_\mu}\right)^{-(\lambda-1)N} \\
&= \sum_{\mu=1}^{s} \binom{\lambda N}{r} p_\mu^r (1-p_\mu)^{N-r} \sum_{i=0}^{\infty} \binom{-(\lambda-1)N}{i} p_\mu^i (1-p_\mu)^{-i} \qquad (13) \\
&= \sum_{i=0}^{\infty} \binom{\lambda N}{r} \binom{-(\lambda-1)N}{i} \sum_{\mu=1}^{s} p_\mu^{r+i} (1-p_\mu)^{N-(r+i)} \\
&= \sum_{i=0}^{\infty} \frac{\binom{\lambda N}{r}\binom{-(\lambda-1)N}{i}}{\binom{N}{r+i}} \mathscr{E}(n_{r+i}). \qquad (14)
\end{aligned}$$

(14) is not rigorously correct, since for $r+i > N$, $\binom{N}{r+i}$ and $\mathscr{E}(n_{r+i})$ both vanish, and the corresponding terms of the series are indeterminate. We notice, however, that if the infinite upper limit for i is replaced by an odd [even] integer, the left-hand side of (13) is greater [less] than the right-hand side†, and the same therefore holds of (14). Thus the partial sums of (14) are alternately greater and less than the left-hand side, and in all practical examples a sufficiently good approximation is reached while $(r+i)$ is still small compared to N. Provided that we use only terms of the series for which $r+i \ll N$ and $i \ll (\lambda-1)N$, we can write

$$\begin{aligned}
\frac{\binom{\lambda N}{r}\binom{-(\lambda-1)N}{i}}{\binom{N}{r+i}} &\doteq \frac{(\lambda N)^r (-(\lambda-1)N)^i (r+i)!}{r!\, i!\, N^{r+i}} \\
&= (-1)^i \lambda^r (\lambda-1)^i \binom{r+i}{r}. \qquad (15)
\end{aligned}$$

Hence

$$\mathscr{E}(n_r(\lambda)) \doteq \lambda^r \sum_{i=0}^{\infty} (-1)^i \binom{r+i}{r} (\lambda-1)^i \mathscr{E}(n_{r+i}), \tag{16}$$

the partial sums erring alternately in excess and defect.

† This follows from the nth Mean Value Theorem applied to $(1+x)^{-(\lambda-1)N}$.

(16) can also be obtained directly by using the Poisson approximation

$$\mathscr{E}(n_r) \doteq \sum_{\mu=1}^{s} e^{-Np_\mu} \frac{(Np_\mu)^r}{r!}. \tag{17}$$

We define an estimate of $\mathscr{E}(n_r(\lambda))$ by

$$\hat{n}_r(\lambda) = \lambda^r \sum_{i=0}^{\infty} (-1)^i \binom{r+i}{r} (\lambda-1)^i n_{r+i}; \tag{18}$$

then (16) gives us $\mathscr{E}(\hat{n}_r(\lambda)) \doteq \mathscr{E}(n_r(\lambda))$.

For the case $r = 0$, we may seem to need to assume the value of s, but this assumption is not really required since we can write

$$\hat{d}(\lambda) = d - \sum_{i=1}^{\infty} (-1)^i (\lambda-1)^i n_i = s - \hat{n}_0(\lambda), \tag{19}$$

so that $\mathscr{E}(\hat{d}(\lambda)) \doteq \mathscr{E}(d(\lambda))$.

We have thus obtained (approximately) unbiased estimates of $\mathscr{E}(n_r(\lambda))$ and $\mathscr{E}(d(\lambda))$ in terms of the observed numbers n_r. We shall almost certainly obtain more accurate estimates if we replace the n_r by smoothed values n_r'; for methods of smoothing see G, §§3, 7, 8.

In the case when the second sample is an enlargement of the basic one, and it is desired to predict the value of $n_r(\lambda)$ *given the basic sample* rather than $\mathscr{E}(n_r(\lambda))$, which we have defined to be the expectation when the whole of the second sample is varied at random, it seems intuitively clear that n_r should not be replaced by n_r' in (18), at any rate when λ is not large, though the later terms probably should be smoothed; for instance, consider the case $\lambda = 1$. We have not attempted a rigorous treatment of this question. The advantage of smoothing is not as great when using formulae like (18) as when using such formulae as G (2′):

$$r^* \doteq (r+1)\, n_{r+1}'/n_r',$$

involving a ratio of the n_r'. Sometimes a ratio is involved surreptitiously, as in G (6), (6′).

An important point in the argument leading to (14) was the expression of $(1-p_\mu)$ as

$$\left(1 + \frac{p_\mu}{1-p_\mu}\right)^{-1},$$

so that expansion led to terms expressible as functions of the $\mathscr{E}(n_r)$. This device can also be used to avoid the approximation made in G (lines 8 and 9 of p. 241) of replacing the expected value of n_{r+m} for a sample of size $N+m$ by its expected value for a sample of size N. The result obtained is (using the notation of G)

$$\mathscr{E}(q_r^m \mid H) = \frac{(r+m)^{(m)}}{(N-r)^{(m)}} \frac{1}{\mathscr{E}(n_r)} \sum_{i=0}^{\infty} \frac{(r+m+i)^{(i)}}{(N-r-m)^{(i)}} \binom{-m}{i} \mathscr{E}(n_{r+m+i}) \tag{20}$$

$$\doteq \frac{(r+m)^{(m)}}{(N-r)^{(m)}} \frac{\mathscr{E}(n_{r+m}) - \dfrac{(r+m+1)\,m}{N-r-m} \mathscr{E}(n_{r+m+1}) + \dots}{\mathscr{E}(n_r)}. \tag{21}$$

The approximations made in G are thus equivalent to replacing $(N-r)^{(m)}$ by N^m and neglecting the terms of the sum after the first; they are reasonable provided $mr \ll N$.

As noted in § 1, we may be particularly interested in the coverage of the second sample, and the number of different species it will contain. By (3) and (16) with $r = 1$, the expected coverage is approximately

$$1-\frac{\mathscr{E}(n_1(\lambda))}{\lambda N}\simeq 1-\frac{1}{N}\sum_{i=0}^{\infty}(-1)^i(i+1)(\lambda-1)^i\mathscr{E}(n_{i+1})$$

$$\simeq 1-\frac{1}{N}[n_1-2(\lambda-1)n_2+3(\lambda-1)^2n_3-\ldots] \tag{22}$$

(or, more accurately, the same formula with n'_r in place of n_r).† The expected number of distinct species represented is (by (19)) approximately

$$d+(\lambda-1)n_1-(\lambda-1)^2n_2+\ldots; \tag{23}$$

i.e. in the case of an enlarged sample, the number of new species expected is approximately

$$(\lambda-1)n_1-(\lambda-1)^2n_2+\ldots. \tag{24}$$

Evidently $n_1, n_2, \ldots$ may be replaced by smoothed values in (23) and (24), but d should be replaced by the smoothed value

$$d' = n'_1+n'_2+\ldots$$

only in the case of an independent second sample. Note that (23) and (24) can be proved directly without assuming s to be finite.

3. Variance of the estimates $\hat{n}_r(\lambda)$

In this section we find an expression for the variance of the estimate $\hat{n}_r(\lambda)$ of $\mathscr{E}(n_r(\lambda))$ defined by (18). This must not be confused with the variance of $n_r(\lambda)$, which can be found from the formulae given in G, § 5. $\hat{n}_r(\lambda)$ is a linear function of the random variables n_{r+i}, and varies accordingly when we take different basic samples; we can find its variance if we know the variances and covariances of the n_r. We therefore start by calculating these. By the method of G, § 5, we find

$$\mathscr{E}(n_r n_s) = \delta_{rs}\mathscr{E}(n_r)+\frac{N!}{r!\,s!\,(N-r-s)!}\sum_{\mu,\nu}^{\mu+\nu}p_\mu^r p_\nu^s(1-p_\mu-p_\nu)^{N-r-s}$$

$$= \delta_{rs}\mathscr{E}(n_r)+\frac{N!}{r!\,s!\,(N-r-s)!}\Big[\sum_\mu\sum_\nu p_\mu^r p_\nu^s(1-p_\mu-p_\nu)^{N-r-s}-\sum_\mu p_\mu^{r+s}(1-2p_\mu)^{N-r-s}\Big], \tag{25}$$

where $\delta_{rr}=1$, $\delta_{rs}=0$ if $r\neq s$.

Now

$$(1-p_\mu-p_\nu)^{N-r-s} = \left[(1-p_\mu)(1-p_\nu)\left(1-\frac{p_\mu}{1-p_\mu}\frac{p_\nu}{1-p_\nu}\right)\right]^{N-r-s}$$

$$= (1-p_\mu)^{N-r}\left(1+\frac{p_\mu}{1-p_\mu}\right)^s(1-p_\nu)^{N-s}\left(1+\frac{p_\nu}{1-p_\nu}\right)^r\left(1-\frac{p_\mu}{1-p_\mu}\frac{p_\nu}{1-p_\nu}\right)^{N-r-s}$$

$$= \sum_{i=0}^{s}\binom{s}{i}p_\mu^i(1-p_\mu)^{N-r-i}\sum_{j=0}^{r}\binom{r}{j}p_\nu^j(1-p_\nu)^{N-s-j}$$

$$\times\sum_{k=0}^{N-r-s}(-1)^k\binom{N-r-s}{k}p_\mu^k(1-p_\mu)^{-k}p_\nu^k(1-p_\nu)^{-k}; \tag{26}$$

† It is correct to replace n_1 by n'_1, even when the second sample is an enlargement of the basic one, because the more accurate formula for the coverage, G (9′), uses n'_1 in place of n_1, so we are interested in $\mathscr{E}(n_1(\lambda))$ rather than the expected value of $n_1(\lambda)$ given the basic sample.

and
$$(1-2p_\mu)^{N-r-s} = \left[(1-p_\mu)\left(1-\frac{p_\mu}{1-p_\mu}\right)\right]^{N-r-s}$$
$$= \sum_{i=0}^{N-r-s}(-1)^i\binom{N-r-s}{i}p_\mu^i(1-p_\mu)^{N-r-s-i}. \tag{27}$$

Substituting from (26) and (27) in (25),

$$\mathscr{E}(n_r n_s) = \delta_{rs}\mathscr{E}(n_r) + \frac{N!}{r!\,s!\,(N-r-s)!}\Bigg[\sum_{i,j,k}(-1)^k\binom{s}{i}\binom{r}{j}\binom{N-r-s}{k}$$
$$\times\Big(\sum_\mu p_\mu^{r+i+k}(1-p_\mu)^{N-r-i-k}\Big)\Big(\sum_\nu p_\nu^{s+j+k}(1-p_\nu)^{N-s-j-k}\Big)$$
$$-\sum_i(-1)^i\binom{N-r-s}{i}\sum_\mu p_\mu^{r+s+i}(1-p_\mu)^{N-r-s-i}\Bigg]$$
$$= \delta_{rs}\mathscr{E}(n_r) + \frac{N!}{r!\,s!\,(N-r-s)!}\Bigg[\sum_{i,j,k}(-1)^k\frac{\binom{s}{i}\binom{r}{j}\binom{N-r-s}{k}}{\binom{N}{r+i+k}\binom{N}{s+j+k}}\mathscr{E}(n_{r+i+k})\mathscr{E}(n_{s+j+k})$$
$$-\sum_i(-1)^i\frac{\binom{N-r-s}{i}}{\binom{N}{r+s+i}}\mathscr{E}(n_{r+s+i})\Bigg]$$
$$= \delta_{rs}\mathscr{E}(n_r) + \sum_{i,j,k}(-1)^k\frac{(N-r-i-k)!\,(N-s-j-k)!\,(r+i+k)!\,(s+j+k)!}{(N-r-s-k)!\,N!\,i!\,j!\,k!\,(s-i)!\,(r-j)!}$$
$$\times\mathscr{E}(n_{r+i+k})\mathscr{E}(n_{s+j+k}) - \sum_i(-1)^i\frac{(r+s+i)!}{r!\,s!\,i!}\mathscr{E}(n_{r+s+i}) \tag{28}$$
$$\doteq \delta_{rs}\mathscr{E}(n_r) + \sum_{i,j,k}(-1)^k\frac{(N-r-i-k)!\,(N-s-j-k)!\,(r+i+k)!\,(s+j+k)!}{(N-r-s-k)!\,N!\,i!\,j!\,k!\,(s-i)!\,(r-j)!}$$
$$\times\mathscr{E}(n_{r+i+k})\mathscr{E}(n_{s+j+k}) - 2^{-r-s}\frac{(r+s)!}{r!\,s!}\mathscr{E}(n_{r+s}(2)), \tag{29}$$

using (8). Provided i, j, k, r, s are all $\ll N$, the coefficient in the first sum is $O((rs/N)^{i+j+k})$; and when $i=j=k=0$, use of Stirling's formula shows that it is $1+O(rs/N)$. Hence, if $rs \ll N$,

$$\operatorname{cov}(n_r, n_s) = \mathscr{E}(n_r n_s) - \mathscr{E}(n_r)\mathscr{E}(n_s)$$
$$\doteq \delta_{rs}\mathscr{E}(n_r) - 2^{-r-s}\binom{r+s}{r}\mathscr{E}(n_{r+s}(2)). \tag{30}$$

Notice that when $r=s$ we have equation (22) of G:

$$V(n_r) \doteq \mathscr{E}(n_r) - 2^{-2r}\binom{2r}{r}\mathscr{E}(n_{2r}(2)), \tag{31}$$

or, expanding the second term by (8),

$$V(n_r) \doteq \mathscr{E}(n_r) - \sum_i(-1)^i\frac{(2r+i)!}{r!\,r!\,i!}\mathscr{E}(n_{2r+i}). \tag{32}$$

For the case $r=0$, since s is constant, we have

$$V(d) = V(n_0) \doteq \mathscr{E}(d(2)) - \mathscr{E}(d) \doteq \mathscr{E}(n_1) - \mathscr{E}(n_2) + \ldots. \tag{33}$$

Using (30), we can now find the variance of $\hat{n}_r(\lambda)$. From the elementary formula for the variance of a linear form:

$$V\Big(\sum_i a_i x_i\Big) = \sum_{i,j} a_i a_j \operatorname{cov}(x_i, x_j),$$

we have

$$
\begin{aligned}
&V(\hat{n}_r(\lambda)) \\
&\quad = \lambda^{2r} \sum_{i,j=0}^{\infty} (-1)^{i+j} (\lambda-1)^{i+j} \binom{r+i}{r} \binom{r+j}{r} \operatorname{cov}(n_{r+i}, n_{r+j}) \\
&\quad \simeq \lambda^{2r} \sum_{i,j=0}^{\infty} (-1)^{i+j} (\lambda-1)^{i+j} \binom{r+i}{r} \binom{r+j}{r} \left[\delta_{ij} \mathscr{E}(n_{r+i}) - 2^{-2r-i-j} \binom{2r+i+j}{r+i} \mathscr{E}(n_{2r+i+j}(2)) \right] \\
&\quad = \lambda^{2r} \left[\sum_{i=0}^{\infty} (\lambda-1)^{2i} \binom{r+i}{r}^2 \mathscr{E}(n_{r+i}) \right. \\
&\qquad \left. - \sum_{l=0}^{\infty} (-1)^l (\lambda-1)^l 2^{-2r-l} \mathscr{E}(n_{2r+l}(2)) \frac{(2r+l)!}{l!\,r!\,r!} \sum_{i,j>0}^{i+j=l} \frac{(i+j)!}{i!\,j!} \right] \\
&\quad = \lambda^{2r} \left[\sum_{i=0}^{\infty} (\lambda-1)^{2i} \binom{r+i}{r}^2 \mathscr{E}(n_{r+i}) - \sum_{l=0}^{\infty} (-1)^l (\lambda-1)^l 2^{-2r} \frac{(2r+l)!}{r!\,r!\,l!} \mathscr{E}(n_{2r+l}(2)) \right] \\
&\quad \simeq \lambda^{2r} \left[\sum_{i=0}^{\infty} (\lambda-1)^{2i} \binom{r+i}{r}^2 \mathscr{E}(n_{r+i}) - \binom{2r}{r} (2\lambda)^{-2r} \mathscr{E}(n_{2r}(2\lambda)) \right], \qquad (34)
\end{aligned}
$$

using (8) again. This derivation of (34) is slightly unsatisfactory, since the second series in the previous formula may give a good approximation to the term by which we replace it only after so many terms have been taken that the approximation made in (30) is no longer valid. It may be possible to postpone making this approximation until a later stage of the calculation, but the algebra would become very heavy. In order to estimate $\mathscr{E}(n_{2r}(2\lambda))$ in calculating (34) for an actual case, it may be possible to use the method of § 2 (probably in conjunction with the summation technique described in § 5), or it may be easier to make a sufficiently accurate guess.

If $\hat{n}_r(\lambda)$ is defined by (18) with n'_{r+i} in place of n_{r+i}, it becomes very difficult to make any estimate of its variance. We can say, however, that so long as we feel that it is worth using smoothed values at all, the variance of the estimate based on them is likely to be considerably less than that given by (34).

4. Variance of $\hat{n}_r(\lambda)$ considered as a prediction of $n_r(\lambda)$

In the last section, we were considering the question: How much may $\hat{n}_r(\lambda)$ be expected to differ from its mean value (which is equal to $\mathscr{E}(n_r(\lambda))$)? A question which may sometimes be more relevant is: How much may $\hat{n}_r(\lambda)$ be expected to differ from the value of $n_r(\lambda)$ obtained in a random second sample? To answer this question, we want to find

$$V(\hat{n}_r(\lambda) - n_r(\lambda)), \qquad (35)$$

which may be called the *variance of* $\hat{n}_r(\lambda)$ *considered as a prediction of the random variable* $n_r(\lambda)$, rather than as an estimate of the parameter $\mathscr{E}(n_r(\lambda))$. It is evidently now necessary to consider separately the cases when the second sample is independent, and when it is an enlargement of the basic sample.

When the second sample is independent, $\hat{n}_r(\lambda)$ and $n_r(\lambda)$ are independent random variables, and so

$$V(\hat{n}_r(\lambda) - n_r(\lambda)) = V(\hat{n}_r(\lambda)) + V(n_r(\lambda)) \qquad (36)$$

which can be calculated by using (34) and the following modification of (31):

$$V(n_r(\lambda)) \simeq \mathscr{E}(n_r(\lambda)) - 2^{-2r} \binom{2r}{r} \mathscr{E}(n_{2r}(2\lambda)). \qquad (37)$$

In the case when the second sample is an enlargement, $\hat{n}_r(\lambda)$ and $n_r(\lambda)$ are correlated, and we have not been able to calculate (35) in this case. It may be expected to be considerably smaller than in the case of independence, at least if λ is not large, since when $\lambda = 1$ we have $\hat{n}_r(1) = n_r = n_r(1)$ and so (35) is reduced to zero.

5. SUMMATION OF THE SERIES OBTAINED IN §§2 AND 3

We consider the case of the general series (18); similar remarks apply to (22), (23), (24) and (32), and to the series arising in calculating (34) and (37). It was pointed out in §1 that the term $(\lambda-1)^i$ in (8) and (18) may cause trouble if $\lambda > 2$. In fact, the series is likely to become 'practically divergent'; i.e. to behave like an infinitely oscillating series up to a point at which we become too uncertain of the value of $\mathscr{E}(n_r)$ to continue with the calculation. This difficulty is illustrated by formula (34) for the variance of $\hat{n}_r(\lambda)$; if $\lambda > 2$, the series

$$\sum_{i=0}^{\infty} \binom{r+i}{r}^2 (\lambda-1)^{2i} \mathscr{E}(n_{r+i})$$

is likely to have a very large sum, unless $\mathscr{E}(n_{r+i})$ decreases extremely fast. It is natural to try to overcome this difficulty by using a method of summation which is known to make some oscillating series converge; a convenient method appears to be that of Euler, with a parameter q, generally called the (E,q) method (Hardy, 1949, pp. 178 ff.). This is to transform the series $\sum_{i=0}^{\infty} a_i$ into $\sum_{j=0}^{\infty} a_j^{(q)}$, where

$$a_j^{(q)} = \frac{1}{(q+1)^{j+1}} \sum_{i=0}^{j} \binom{j}{i} q^{j-i} a_i \tag{38}$$

$$= (-1)^j \left(\frac{q}{q+1}\right)^{j+1} \Delta_0^j \left[(-1)^i \frac{a_i}{q^{i+1}}\right], \tag{39}$$

the forward difference symbol Δ_r^j being defined inductively by

$$\Delta_r^1 a_i = a_{r+1} - a_r, \quad \Delta_r^j = \Delta_r^1 \Delta_i^{j-1}. \tag{40}$$

(The form (39) is given by Bromwich (1926), pp. 62–6, for the case $q = 1$. It leads to a convenient method of setting out the work in a practical example, which will be illustrated in Example (i) of the next section.) If $\sum_i a_i$ converges, then $\sum_j a_j^{(q)}$ converges to the same sum, for any $q \geqslant 0$; if $\sum_j a_j^{(q')}$ converges, then $\sum_j a_j^{(q)}$ converges to the same sum for all $q \geqslant q'$ (Hardy, 1949, Theorems 117 and 118).

In practical examples, n_{r+i} generally decreases slowly after the first few terms, and we are usually interested in small values of r, so that $\binom{r+i}{r}$ increases slowly. Under these circumstances, the series (18) is, after the first few terms, nearly a G.P. with ratio $-(\lambda-1)$. Now, if we apply the (E,q) method to such a G.P., say

$$a_i = (-1)^i (\lambda-1)^i a_0, \tag{41}$$

we obtain

$$a_j^{(q)} = \frac{a_0}{(q+1)^{j+1}} \sum_{i=0}^{j} \binom{j}{i} q^{j-i} (-(\lambda-1))^i$$

$$= \frac{a_0}{(q+1)^{j+1}} (q-(\lambda-1))^j$$

$$= \frac{a_0}{q+1} \left(\frac{q-(\lambda-1)}{q+1}\right)^j, \tag{42}$$

i.e. the transformed series is a G.P. with ratio $\frac{q-(\lambda-1)}{q+1}$. Clearly the best value of q to select is $\lambda-1$, which reduces all but the first term of the transformed series to zero.† If the n_{r+t} decrease fairly rapidly, we may get better results by choosing q somewhat smaller. (This is the case in Example (i) of § 6, where $\lambda = 5$, but we take $q = 2$.) When $r = 0$ or 1, and if $n_1 \gg n_2$, it may be worth taking out the first term of the series, and applying the summation process to the remainder; the reason for this can be seen by considering such a series as

$$1 - \tfrac{1}{8} + \tfrac{1}{4} - \tfrac{1}{2} + \ldots. \tag{43}$$

If we apply the $(E, 2)$ method directly, we get

$$0{\cdot}333 + 0{\cdot}208 + 0{\cdot}139 + 0{\cdot}093 + \ldots, \tag{44}$$

while if we take out the first term and apply the $(E, 2)$ method to the remainder, we get

$$1{\cdot}000 - 0{\cdot}042 + 0{\cdot}000 + 0{\cdot}000 + \ldots, \tag{45}$$

which is evidently better.

When we have chosen a method of summation, and selected a partial sum of the transformed series as probably giving a sufficiently good approximation to the final sum, we can express this partial sum as a linear combination of the n_r, and deduce its variance, as in § 3.‡ But there is now a new source of error, namely, the omission of the rest of the transformed series. We have not been able to find a useful form of error term for this remainder (corresponding to the statement that alternate partial sums of (8) err in excess and defect); failing such an error term, our results must be used with caution when it is necessary to apply the summation process. If the n_r's decrease slowly and q is taken to be slightly smaller than $\lambda-1$, the transformed series will generally have terms alternating in sign (cf. equation (42)); it might then be hoped that the partial sums err alternately in excess and defect, but it does not seem to be possible to lay down any simple general conditions under which this is the case.

Some of the methods described by Shanks (1955) also seem to be very well suited to our case; they have the property of summing perfectly any series which is geometric from some point onwards, so that the difficulty caused by an excessively large first term, noted above, does not arise. Given the series $\sum_{n=0}^{\infty} a_n$ we define a sequence (not a new series) by

$$B_n = \sum_{r=0}^{n-1} a_r + \frac{a_n^2}{a_{n+1} - a_n} \quad (n = 1, 2, 3, \ldots);$$

repetition of the process gives a sequence C_n, and so on. The e_1 method consists of considering the sequence B_n (in place of the sequence of partial sums $A_n = \sum_{r=0}^{n} a_r$), the e_1^2 method, of considering the sequence C_n, and so on; the $\tilde{e}_1$ method consists of considering the sequence $A_0, B_1, C_2, \ldots$. For an example, see § 6, Example (i).

† This statement may appear to conflict with the remark of Hardy (1949), p. 180, that 'as q increases, the (E, q) methods form a scale of increasing strength'. But here 'strength' refers only to whether we obtain a convergent series or not: if we choose q unnecessarily large, we shall certainly obtain a convergent series, but it will converge very slowly.

‡ This remark applies to any method of summation by a linear transformation of the series, e.g. Cesàro means, the composite $(E, q; C, k)$ method, any Hausdorff means, Hölder means, Hutton's method, any Nörlund means, or quasi-Hausdorff transformations; for references to all these methods, see Hardy (1949), p. 392. It does not apply to the non-linear methods of Shanks (1955), mentioned below.

6. EXAMPLES

The first example is an artificial one designed to test the efficacy of the methods described above, especially the summation methods of §5. The second and third examples illustrate the practical applications, but enlarged samples are not available for verifying the estimates.

Example (*i*). *Sample of words from 'Our Mutual Friend' by Charles Dickens.* The following samples were taken:

A, of 1000 words, the last words of lines on pages $\equiv 5 \bmod 25$,

B, of 2000 words, the last words of lines on pages $\equiv 10$ or $20 \bmod 25$,

C, of 2000 words, the last words of lines on pages $\equiv 15$ or $25 \bmod 25$,

the sampling in each case being carried as far as required to make up the prescribed number of words. Our original intention was to use A as the basic sample ($N = 1000$) and to calculate

Table 1

Sample A; $N = 1000$			Sample A; $N = 1000$		
r	n_r	n'_r	r	n_r	n'_r
1	404	404	6	3	—
2	57	64	7	0	—
3	24	25	8	3	—
4	16	12·2	$\geqslant 9$	15	—
5	6	6·2			
$d = 528$					

values of $\hat{d}(\lambda)$ and $\hat{n}_1(\lambda)$ given by (19) and (18) for $\lambda = 2, 3, 4, 5$, which could be checked against the values of $d(\lambda)$ and $n_1(\lambda)$ actually obtained from the samples B, $A+B$, $B+C$, $A+B+C$. The results, however, showed a systematic and, for $\hat{d}(2)$, significant difference between the prediction and the observed result. Working back from sample B with $\lambda = \frac{1}{2}$, it appeared that sample A had n_1 considerably too small. We believe that this is due to the fact that the method of sampling used was not sufficiently random; an uncommon word is likely to occur several times on the same page, where a particular topic is discussed, and such a word is therefore less likely to occur just once in a sample selected as described than in a random sample of the same size.†

The results for larger values of λ were, however, not much less accurate than those for $\lambda = 2$, and we give the calculation of $\hat{d}(5)$ as an example of the use of the (E, q) method of summation described in §5. Table 1 shows the data; the n'_r were obtained by graphical smoothing of $\sqrt{n_r}$. Our formula (19) gives us

$$\hat{d}(5) = 528 + (4.404 - 4^2.64 + 4^3.25 - 4^4.12{\cdot}2 + 4^5.6{\cdot}2 - \ldots). \tag{46}$$

† Consider the extreme case when p. 1 reads 'one one one...', p. 2 'Two two two...', and so on.

To transform the bracketed series we form the difference table suggested by (39) (q has been chosen as 2, so as to make the differences small):

$\frac{1}{2}.4.404$	$=808$			
		-552		
$(\frac{1}{2})^2.4^2.64$	$=256$		496	
		-56		-445
$(\frac{1}{2})^3.4^3.25$	$=200$		51		402
		-5		-43
$(\frac{1}{2})^4.4^4.12{\cdot}2$	$\simeq 195$		8	
		3		-488
$(\frac{1}{2})^5.4^5.6{\cdot}2$	$\simeq 198$		555	
		-610		

(We apply the usual check, that the sum of each column is equal to the difference between the top and bottom of the one before.) The transformed series, by (39), is

$$\tfrac{2}{3}.808+(\tfrac{2}{3})^2.552+(\tfrac{2}{3})^3.496+(\tfrac{2}{3})^4.445+(\tfrac{2}{3})^5.402\ldots\simeq 538+245+147+88+53\ldots. \quad (47)$$

The last few terms of (47) are approximately a geometric series with ratio 0·6; the sum of the remaining terms should therefore be approximately†

$$53\times\frac{0{\cdot}6}{1-0{\cdot}6}\simeq 79,$$

making a total of 1150; hence

$$\hat{d}(5) = 528+1150 = 1678. \quad (48)$$

Applying the methods of Shanks (1955), described at the end of §5, we get Table 2. Although the transformed sequences are rather short, it looks as if $C_2 = 1155$ is a good approximation to the limit, giving $\hat{d}(5) = 1683$. In fact, for the whole sample $A+B+C$, $d(5) = 1832$.

Table 2

n	A_n	B_n	C_n
0	1616		
1	592	1216	
2	2192	1134	1155
3	−931	1162	
4	5418		

In this example, we have been slightly handicapped by having so few terms of the series available; when using the (E, k) method, this renders the remainder somewhat uncertain, and prevents us from omitting the first term from the summation process, as was suggested in §5. Because of this difficulty and the apparent non-randomness noted above, we use sample B ($N = 2000$) as the basic sample for a more comprehensive test of our methods, although we can then verify the results only up to $\lambda = 2{\cdot}5$. (The 'second samples' for $\lambda = 1{\cdot}5$, 2·0, 2·5 are $A+B$, $B+C$, $A+B+C$ respectively, and are thus all enlargements of the basic sample.) Table 3 gives the data for this basic sample; the n'_r were produced

† This is, as Shanks (1955) points out, equivalent to applying the e_1 method to sum (47).

by smoothing $\sqrt{n_r}$ graphically by the use of French curves, the n_r'', independently, by smoothing by eye: $\hat{d}(\lambda)$, $\hat{n}_1(\lambda)$, and the estimated percentage coverage, $100\,(1-\hat{n}_1(\lambda)/\lambda N)$, were calculated for $\lambda = 1{\cdot}5,\ 2{\cdot}0,\ 2{\cdot}5$, using the three sets of values. The summation process was used only in the case $\lambda = 2{\cdot}5$, with $q = 1$. Table 4 shows the three sets of estimates and the actual results found in the enlarged samples; standard deviations are given where applicable, calculated from (31), (33), and (34). It will be noticed that in this case little or nothing was gained when smoothed values were used; but it would probably be essential to use smoothed values when working with larger values of λ.

Table 3

Sample B; $N = 2000$			
r	n_r	n_r'	n_r''
1	729	729	729
2	108	96	110
3	33	38	38
4	23	21	19
5	17	14	13
6	7	9	9
7	5	7	6
8	3	3·2	4
$\geqslant 9$	30	—	—
$d = 955$			

Example (*ii*). *Captures of Macrolepidoptera in a light-trap at Rothamsted.* (Quoted as example (i) in G, §8 from Williams's data in Corbet, Fisher & Williams (1943).) $N = 15609$, $d = 240$. Table 5 shows the small values of r. n_r' is n_r''' of the example in G, obtained by smoothing $\sum_{t=1}^{r} tn_t$. n_r'' is Fisher's analytic smoothing, given by H_3 of G with parameter $\beta = 40{\cdot}2$. Now H_3 is a hypothesis defining the distribution of the population frequencies $\{p_\mu\}$, and it implies that

$$\mathscr{E}(n_r(\lambda)) \simeq \frac{\beta}{r}\left(\frac{N\lambda}{N\lambda+\beta}\right)^r. \tag{49}$$

(G, (63)), and

$$\mathscr{E}(d(\lambda)) = \beta \log_e\left(\frac{N\lambda}{\beta}+1\right) \tag{50}$$

(G, (67)). Since $N \gg \beta$, we see that H_3 implies

$$\mathscr{E}(n_1(\lambda)) \simeq \mathscr{E}(n_1) \tag{51}$$

and

$$\mathscr{E}(d(\lambda)) \simeq d + \beta \log_e \lambda. \tag{52}$$

Putting $\lambda = 2$ we see that doubling the sample will approximately halve the proportion of the population not represented (by (51) and (3)) and increase the number of distinct species observed by approximately $\beta \log_e 2 = 27{\cdot}9$. (The latter fact was noted by Williams in Corbet *et al.* (1943), p. 51.)

Table 4

Estimates of $d(\lambda)$	$\lambda = 1{\cdot}5$	$\lambda = 2{\cdot}0$	$\lambda = 2{\cdot}5$
Using n_r	1296 ± 34	1599 ± 50	1872
Using n_r'	1299	1613	1909
Using n_r''	1296	1601	1890
Actual results	1303 ± 29	1551 ± 31	1832 ± 34

Estimates of $n_1(\lambda)$	$\lambda = 1{\cdot}5$	$\lambda = 2{\cdot}0$	$\lambda = 2{\cdot}5$
Using n_r	958 ± 42	1172†	1322
Using n_r'	981	1238	1435
Using n_r''	961	1168	1350
Actual results	983 ± 29	1116 ± 31	1308 ± 32

Estimates of % coverage	$\lambda = 1{\cdot}5$	$\lambda = 2{\cdot}0$	$\lambda = 2{\cdot}5$
Using n_r	$68{\cdot}1 \pm 1{\cdot}4$	70·7†	73·6
Using n_r'	67·3	69·0	71·3
Using n_r''	68·0	70·8	73·0
Actual results	$67{\cdot}2 \pm 1{\cdot}0$	$72{\cdot}1 \pm 0{\cdot}8$	$73{\cdot}8 \pm 0{\cdot}6$

† The S.D. is not given in this case because the sum of the series was not taken to infinity but estimated after eight terms as lying midway between the last partial sums; that is, in effect, Hutton's method (Hu, 1) was applied to sum the series (Hardy, 1949, pp. 21–2). (34) would give a very large variance, most of which arises from terms after the eighth.

Table 5

r	n_r	n_r'	n_r''
1	35	35·0	40·1
2	11	22·5	20·0
3	15	16·3	13·3
4	14	12·3	10·0
5	10	9·7	7·9
6	11	7·7	6·6
7	5	6·0	5·6

The present example is not a good one to which to apply our distribution-free† methods, since even n_1 is rather small; however, we shall obtain the corresponding results for comparison, using the smoothed values n'_r. By (18),

$$\begin{aligned}\hat{n}_1(2) &\simeq 2(35{\cdot}0 - 45{\cdot}0 + 48{\cdot}9 - 49{\cdot}2 + 48{\cdot}5 - 46{\cdot}2 + 42{\cdot}0 - \ldots) \\ &= 2(17{\cdot}5 - 2{\cdot}5 - 0{\cdot}8 - 0{\cdot}2 + 0{\cdot}0 + 0{\cdot}1 + 0{\cdot}1 + \ldots)\end{aligned}$$

(transforming the series by $(E, 1)$)

$$= 28{\cdot}4,$$

whence (using (3)) we predict that the proportion of the population not covered should decrease from

$$\frac{35{\cdot}0}{15609} = 0{\cdot}22\,\%$$

to

$$\frac{28{\cdot}4}{31218} = 0{\cdot}09\,\%,$$

whereas H_3 implied that it was halved. By (24), the expected number of new species is approximately

$$\begin{aligned}&35{\cdot}0 - 22{\cdot}5 + 16{\cdot}3 - 12{\cdot}3 + 9{\cdot}7 - 7{\cdot}7 + 6{\cdot}0 - \ldots \\ &\quad = 17{\cdot}5 + 3{\cdot}1 + 0{\cdot}8 + 0{\cdot}3 + 0{\cdot}1 + 0{\cdot}1 + 0{\cdot}0 + \ldots \quad \text{(by } (E, 1)\text{)} \\ &\quad = 21{\cdot}9,\end{aligned}$$

whereas H_3 implied that 27·9 were expected. Finally, in order to estimate the S.D. of n_1 we calculate from (18)

$$\begin{aligned}\hat{n}_2(2) &\simeq 4(22{\cdot}5 - 48{\cdot}9 + 73{\cdot}8 - 97{\cdot}0 + 115{\cdot}5 - 126{\cdot}0 + \ldots) \\ &= 4(11{\cdot}25 - 6{\cdot}60 - 0{\cdot}19 + 0{\cdot}01 - 0{\cdot}09 - 0{\cdot}04 + \ldots) \quad \text{(by } (E, 1)\text{)} \\ &= 17{\cdot}4,\end{aligned}$$

so, by (31),

$$V(n_1) \simeq 35 - \tfrac{1}{4}.2.17{\cdot}4 = 26{\cdot}3,$$

giving n_1 a S.D. of about 5·1, whereas H_3 (by (65) of G) gave 5·5.

We note that our distribution-free estimates of $n_1(2)$, $n_2(2)$, $d(2)$ are all less than the values implied by H_3. This is what one would expect when the sample size, N, is large enough for the finiteness of s (the total number of species in the population) to conflict with the prediction of H_3 that $s = \infty$; but the effect may well be accidental.

In this example, we can also make a rough comparison between the variances of the estimate of $n_1(\lambda)$ deduced from H_3 and that given by our distribution-free method. The former estimate is almost exactly β, and its variance is therefore approximately equal to the sampling variance of β for this example, given by Fisher (Corbet *et al.* (1943), p. 56) as 1·13 (S.D. of 1·06); this is independent of λ. Our $\hat{n}_1(\lambda)$, on the other hand, has a very large variance for $\lambda \geqslant 2$ (by (34)), and even for $\lambda = 1$, since $\hat{n}_1(1) = n_1$, its variance, as we saw above, is about 26·3 (S.D. 5·1). However, it must be remembered, first, that H_3 is certainly not exactly true (since in fact $s < \infty$),‡ so that the estimate deduced from it is subject to an unknown additional error, and secondly, that we may hope to reduce the variance of $\hat{n}_r(\lambda)$ considerably by using carefully smoothed values and summation methods.

† I.e. independent of any particular assumption about the distribution of the p_μ, in contrast to the above argument which assumes H_3.

‡ Not even the truncated form, H_5 of G, can be exactly true, as was shown in G, p. 257.

In general, if a fairly simple hypothesis H on the p_μ (e.g. any of H_1 to H_6 of G) gives a good fit for the n_r, we should prefer to deduce $\mathscr{E}(n_r(\lambda))$ and $\mathscr{E}(d(\lambda))$ from H, rather than use the distribution-free estimates (18) and (19); but such extrapolation should be made with caution, and the distribution-free methods may give a useful indication of the error to be expected if H is false.

Example (*iii*). *Sample of nouns in Macaulay's essay on Bacon.* (From Yule (1944), Table 44, p. 163; quoted in G as example (iii), p. 260.) $N = 8045$, $d = 2048$.

Table 6

r	n_r	n'_r	r	n_r	n'_r
1	990	1024	11	24	15·5
2	367	341	12	19	13·1
3	173	170	13	10	11·3
4	112	102	14	10	9·7
5	72	68	15	13	8·5
6	47	49	16–20	31	30·5
7	41	35·5	21–30	31	31·5
8	31	28·5	31–50	19	25·9
9	34	22·7	51–100	6	19·9
10	17	18·4	101–∞	1	20·3

(As in the tables in G, n_r and n'_r have been summed where values of r are grouped.)

Here n'_r ($= n''_r$ of G) is the analytic smoothing

$$n'_r = \frac{2048}{r(r+1)}. \tag{53}$$

(H_9 of G; notice that this is *not* an explicit hypothesis on the p_μ, and that it gives a good fit only for $r \leqslant 30$.) (53) is so simple in form that we can carry through all our calculations analytically. Again we consider doubling the sample ($\lambda = 2$); by (18),

$$\begin{aligned}\hat{n}_1(2) &= 2\sum_{i=0}^{\infty}(-1)^i(i+1)\frac{2048}{(i+1)(i+2)}\\ &= 2.2048\sum_{i=0}^{\infty}\frac{(-1)^i}{i+2}\\ &= 2.2048(1-\log_e 2)\\ &\doteq 1260;\end{aligned}$$

and

$$\begin{aligned}\hat{n}_2(2) &= 4\sum_{i=0}^{\infty}(-1)^i\frac{(i+1)(i+2)}{2}\frac{2048}{(i+2)(i+3)}\\ &= 2.2048\sum_{i=0}^{\infty}(-1)^i\frac{i+1}{i+3}\\ &= 2.2048\left[\sum_{i=0}^{\infty}(-1)^i - 2\sum_{i=0}^{\infty}\frac{(-1)^i}{i+3}\right].\end{aligned}$$

The first series is summed by any standard method to $\frac{1}{2}$, and the second is equal to $\log_e 2 - \frac{1}{2}$; hence

$$\begin{aligned}\hat{n}_2(2) &= 2.2048(\tfrac{3}{2} - 2\log_e 2)\\ &\doteq 465.\end{aligned}$$

Finally, by (19),

$$\begin{aligned}\hat{d}(2) &= 2048 - \sum_{i=1}^{\infty} (-1)^i \frac{2048}{i(i+1)} \\ &= 2048\left[1 - \sum_{i=1}^{\infty}\left(\frac{(-1)^i}{i} + \frac{(-1)^{i+1}}{i+1}\right)\right] \\ &= 2048.2\sum_{i=1}^{\infty}\frac{(-1)^{i+1}}{i} \\ &= 2048.2\log 2 \\ &\doteq 2840.\end{aligned}$$

Notice that, since the n'_r give a good fit only for $r \leqslant 30$, the justification for substituting them in infinite series rests on the following argument:

(i) the partial sum of the series† with the true $\mathscr{E}(n_r)$ down to the term containing $\mathscr{E}(n_{30})$ is a good approximation to its infinite sum;

(ii) the same is true for the series† with the n'_r;

(iii) the n'_r are good approximations to the $\mathscr{E}(n_r)$ for $r \leqslant 30$, so that the partial sums mentioned are nearly equal.

(Compare the argument justifying evaluation of integrals by the saddle-point method.) (i)–(iii) should be borne in mind whenever an analytic smoothing is used in this way, or even when it is used to give values of n'_r which are treated numerically; otherwise there is some risk of obtaining an apparently satisfactory convergence which is in fact spurious. When possible, it would probably be advisable in such cases to try a graphical smoothing as well: the reader might like to try the smoothing n'_r of G, using the $(E, 1)$ or $\mathscr{E}_1$ method of summation.

We have now sufficient data to derive the result which was quoted in §1. By (3), the proportion of the population not represented in the 2048 nouns of the basic sample is about‡

$$\tfrac{1024}{8046} \doteq 12{\cdot}7\ \%;$$

by (7) of G the proportion not represented in the 2840–1260 = 1580 nouns occurring twice or more in the doubled sample will be about

$$\frac{1260 + 2.465}{16090} \doteq 13{\cdot}6\ \%;$$

by (2) of G, the average frequency of the 1260 nouns occurring once only in the doubled sample will be about

$$\frac{2.465}{1260.16090} \doteq 0{\cdot}0046\ \%.$$

Hence, if we add a random selection of

$$\frac{13{\cdot}6 - 12{\cdot}7}{0{\cdot}0046} \doteq 200$$

of the nouns occurring once only in the doubled sample to all those occurring twice or more. we will have a list of about 1780 nouns covering approximately the same proportion of the population as the 2048 nouns of the basic sample.

† Or, if summation methods are used (as in calculating $\hat{n}_1(2)$ above), the sum of the transformed series.

‡ The figure of 12·3 % given in G was based on the unsmoothed values.

APPENDIX

Conditions for the lemma of § 1

Although this lemma was not actually used in our argument, we give here, for any reader who may be interested, two fairly general sets of conditions under which it holds.

If $a_r \geqslant 0$ *for all* r, *and finite numbers* b_i *are defined by* (4), *then* (5) *holds if and only if*

$$\frac{b_{r+i}}{i!} \to 0 \quad \text{as} \quad i \to \infty. \tag{54}$$

Proof. Write

$$R(n,r) = \frac{1}{r!}\sum_{i=0}^{n}\frac{(-1)^i b_{r+i}}{i!} - a_r,$$

so that (5) holds if and only if $R(n,r) \to 0$ as $n \to \infty$. Now, for all $n > r$,

$$\begin{aligned} R(n,r) &= \frac{1}{r!}\sum_{i=0}^{n}\frac{(-1)^i}{i!}\sum_{s=0}^{\infty} s^{(r+i)}a_s - a_r \\ &= \sum_{s=0}^{\infty}\binom{s}{r}a_s\sum_{i=0}^{n}(-1)^i\binom{s-r}{i} - a_r \\ &= \sum_{s=0}^{\infty}(-1)^n\binom{s}{r}\binom{s-r-1}{n}a_s - a_r, \end{aligned}$$

using the definition $\binom{a}{b} = \frac{a^{(b)}}{b!}$, even if a is negative, together with the well-known identity

$$\binom{s-r}{i} = \binom{s-r-1}{i-1} + \binom{s-r-1}{i}.$$

Putting $s = r$ gives a term $+a_r$, and all other terms with $s < n+r+1$ vanish; hence

$$\begin{aligned} R(n,r) &= (-1)^n\sum_{s=n+r+1}^{\infty}\binom{s}{r}\binom{s-r-1}{n}a_s \\ &= (-1)^n\sum_{s=n+r+1}^{\infty}\frac{s-r-n}{s-r}\frac{s^{(n+r)}}{n!\,r!}a_s. \end{aligned} \tag{55}$$

Now, since $a_s \geqslant 0$ for all s,

$$\begin{aligned} |R(n,r)| &\leqslant \frac{1}{n!\,r!}\sum_{s=0}^{\infty}s^{(n+r)}a_s \\ &= \frac{1}{r!}\frac{b_{r+n}}{n!}; \end{aligned}$$

and the sufficiency of (54) follows. The necessity is trivial, since if (54) does not hold the right-hand side of (5) cannot converge.

If $a_r = O(x^r)$, $0 \leqslant x < \frac{1}{2}$, *then* (5) *holds for all* r; *further*, (5) *does not hold if* $a_r = 2^{-r}$, *so this result cannot be improved by extending the range of* x.

Proof. (i) We may assume without loss of generality that $|a_r| \leqslant x^r$. Then it follows from (55) above that

$$\begin{aligned} |R(n,r)| &\leqslant \sum_{s=n+r}^{\infty}\frac{s^{(n+r)}}{n!\,r!}x^s \\ &= \frac{(n+r)!\,x^{n+r}}{n!\,r!}\sum_{t=0}^{\infty}\binom{-n-r-1}{t}(-x)^t \\ &= \frac{1}{r!\,x}(n+r)^{(r)}\left(\frac{x}{1-x}\right)^{n+r+1} \\ &\to 0 \text{ as } n \to \infty \text{ provided } x < \tfrac{1}{2}; \end{aligned}$$

hence (5) holds for all r.

(ii) Taking $a_r = 2^{-r}$, we have

$$\begin{aligned} b_i &= \sum_{r=0}^{\infty}r^{(i)}(\tfrac{1}{2})^r \\ &= 2.i!, \end{aligned}$$

summing the series as in (i), and it is clear that the right-hand side of (5) is not convergent for any r.

REFERENCES

BROMWICH, T. J. I'A. (1926). *An Introduction to the Theory of Infinite Series*, 2nd ed. Cambridge University Press.

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.

DICKENS, CHARLES. *Our Mutual Friend.* London: Thomas Nelson. (First published in 1864–5.)

GOOD, I. J. (1953) (described in text as G). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–64.

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

SHANKS, D. (1955). Nonlinear transformations of divergent and slowly convergent series. *J. Math. Phys.* **34**, 1–42.

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