A. The precision of the estimation of $\Sigma(p^2)$

Precision of the estimation of $S(p^2)$

If a_i is any frequency observed in a multinomial distribution and p_i the probability of any observation falling in that class, then it is well known and easily demonstrated that the expectations of the following simple functions of the observations are as shown below.

Function	Expectation			
$\Sigma\,(a)$	$n\,\Sigma(p)$	$=$	n	(1)
$\Sigma\{a(a-1)\}$	$n(n-1)\Sigma(p^2)$	$=$	$n(n-1)s_2$	
$\Sigma\{a(a-1)(a-2)\}$	$n(n-1)(n-2)\Sigma(p^3)$	$=$	$n(n-1)(n-2)s_3$	

and so on.

In the absence of knowledge of the number, or true frequencies of the distinguishable classes, an unbiased estimate can be made of the parameter

$$\alpha = \Sigma(p^2) = s_2, \qquad (2)$$

by using the observable statistic~~s~~

$$\Sigma\{a(a-1)\} \div n(n-1)\ ; \qquad (3)$$

it may be required to assign a sampling variance ~~of~~ to this estimate.

Now, the mean value of the square

$$\Sigma^2\{a(a-1)\} \qquad (4)$$

may be expressed as that of the sum for all classes of

1 - 5 + 6
4 - 8
+ 2

$$\left.\begin{aligned} &a^2(a-1)^2 \\ &= a(a-1)(a-2)(a-3) + 4a(a-1)(a-2) + 2a(a-1)\end{aligned}\right\} \qquad (5)$$

of which the mean is

$$n(n-1)\left\{2s_2 + 4(n-2)s_3 + (n-2)(n-3)s_4\right\}\,2, \qquad (6)$$

and that of the sum of products for all pairs of classes

$$2a(a-1)\,b\,(b-1), \qquad (7)$$

of which the mean is

$$n(n-1)(n-2)(n-3)\,\Sigma\Sigma\,(2p^2q^2), \qquad (8)$$

and which together with the ~~last~~ term in s_4 gives

$$n(n-1)(n-2)(n-3)s_2^2 \qquad (9)$$

Consequently, the sampling variance of the estimate~~s~~ of the parameter α is

$$\cancel{n(n-1}) \left\{2s_2 + 4(n-2)s_3 - (4n-6)s_2^2\right\} / n(n-1) \qquad (10)$$

which is not expressible wholly in terms of s_2, but requires also a knowledge of the sum of the third powers, s_3. In general, we may expect s_3 to lie between α and α^2, but the limits of uncertainty are generally rather wide~~r~~, $\left\{ \frac{~~2s_2(1-s_2)~~}{2\alpha(1-\alpha)}, \; \frac{~~(4n-6)s_2(1-s_2)~~}{(4n-6)\alpha(1-\alpha)} \right\}$.

In the case that the frequencies are those of individual alleles in a self-sterility series, the probabilities p have a distribution, which, for statistical equilibrium, is known in terms of the parameter α, and the population size, N.

The distribution is

$$\frac{1}{p} \cdot \frac{1}{(N-1)!} x^{N-1} e^{-x} dx, \qquad (11)$$

when x stands for

$$\frac{N(1-2p)}{1-2\alpha} \qquad (12)$$

with the understanding that when p becomes small the continuous Eulerian form is replaced by a terminating series with

$$p = {}^{r}/_{2N} \qquad (13)$$

r taking integral values down to unity.

For this distribution it is easily found that

$$\left.\begin{aligned} &\Sigma(p) = 1 \\ &\Sigma p(1-2p) = 1 - 2\alpha, \qquad \Sigma(p^2) = \alpha \\ &\Sigma p(1-2p)^2 = \left(1+\frac{1}{N}\right)(1-2\alpha)^2, \quad \Sigma(p^3) = \alpha^2 + \frac{1}{4N}(1-2\alpha)^2 \end{aligned}\right\} \qquad (14)$$

Using this value in the estimated variance of an estimate of α derived from n observations, we have

$$\left.\begin{aligned} V(\alpha) &= \frac{1}{n(n-1)} \left\{ 2s_2 + 4(n-2)s_3 - (4n-6)s_2^2 \right\} \\ &= \frac{1}{n(n-1)} \left\{ \begin{aligned} &2\alpha + 4(n-2)\alpha^2 + \frac{n-2}{N}(1-2\alpha)^2 \\ &- (4n-6)\alpha^2 \end{aligned} \right\} \end{aligned}\right\} \qquad (15)$$

$$= \frac{2\alpha(1-\alpha)}{n(n-1)} + \frac{n-2}{n(n-1)N}(1-2\alpha)^2 \Bigg) \qquad (15)$$

or, for large populations, supposedly nearly in equilibrium

$$V(\alpha) = \frac{2\alpha(1-\alpha)}{n(n-1)} \quad ; \qquad (16)$$

the precision is therefore about that of a frequency based on $\frac{1}{2}n(n-1)$ observations.

For example, in 25 plants of a population of red clover, Bateman found 36 alleles occurring once only, while 7 occurred twice. Treating them as 50 independent observations

$$\Sigma\ a(a-1) = 14 \qquad (17)$$
$$\cancel{\Sigma}\ n(n-1) = 2450$$

giving the estimate of α

$$1/175 = \cdot 00571428. \qquad (18)$$

The population size N may be presumed to be large compared with

$$\tfrac{1}{2} \cdot 48 \cdot \overset{172}{\cancel{135}} = \overset{4128}{\cancel{3240}}, \qquad (19)$$

so the second term in the expression for the variance is ignored; the variance is then

$$V(\alpha) = \frac{\alpha(1-\alpha)}{1225} = \cdot 0^5 463807 \qquad (20)$$

and the standard error of random sampling is

$$\cdot 002154$$

For 7 observations out of 1225, the equivalent binomial observation, Stevens' Table of approximate fiducial limits gives

$$2.81 \text{ and } 14.38,$$

or odds of 19 to 1 that α lies between the limits

$$\cdot 00229 \text{ and } \cdot 01174.$$

The number of alleles should then lie between

$$85 \text{ and } 436$$

It would have made little difference if, as might be preferred to allow for the fact that in a sample of heterozygotes the observed frequencies are not quite independent the estimate had been taken to be

$$14/50 \times 48$$

instead of $14/50 \times 49$.