

A PROBLEM INVOLVING THE LEXIS THEORY OF DISPERSION

BY WALTER A. HENDRICKS

The attention of the author was recently directed to a study of the hatchability of chicken eggs at the U. S. Animal Husbandry Experiment Station, Beltsville, Maryland. It was necessary to find the average hatchability of the fertile eggs incubated for each of a number of lots of birds and the corresponding standard errors of those averages.

It was very apparent that some methods for computing such values, in common use at the present time, do not give satisfactory results. This is due to the fact that the fertile eggs produced by different birds vary considerably with respect to hatchability as well as with respect to number of eggs available for incubation. It seems reasonable to suppose that the variability in hatchability of a number of fertile eggs, produced by a given number of birds, should obey the Lexis law of dispersion. This supposition is based on two hypotheses:

(a) The probability that a fertile egg will hatch is constant for all fertile eggs produced by the same bird during the time interval under consideration.

(b) The probability that a fertile egg will hatch varies from bird to bird.

The reader familiar with the principles of genetics may question the validity of the first of these hypotheses. The probability that a fertile egg will hatch is largely governed by the genes carried by the chromosomes of the ovum of the hen and the sperm of the male bird which fertilized that ovum. The kinds of genes carried by various ova and spermatozoa are not necessarily the same, even when those ova and spermatozoa are produced by the same female and male birds, respectively. However, if we have a sample of a number of fertile eggs produced by the same hen, we are justified in assuming that the proportion of those eggs which will hatch is constant, except for sampling fluctuations, when successive samples of fertile eggs produced by the given hen are incubated, provided, of course, that the eggs in the successive samples were fertilized by the same male bird or birds. The limit approached by the proportion of fertile eggs which hatch as the number of fertile eggs produced by the given hen becomes infinitely large may be defined as the probability that a fertile egg produced by that hen will hatch. It will be recognized that this definition is based on purely academic considerations, since there are physical limitations to the number of fertile eggs which a hen can produce in a given period of time. Hypotheses (a) and (b) are to be interpreted in the light of this definition of the probability that a fertile egg produced by a given bird will hatch.

Let s_1, s_2, \dots, s_n represent the numbers of fertile eggs produced by n birds during a period of time and let f_1, f_2, \dots, f_n , respectively, represent the numbers

of chicks obtained from those eggs when the eggs are incubated. Let $p_k = \frac{f_k}{s_k}$ represent the hatchability of the fertile eggs produced by the k^{th} bird.

The squared standard error of p_k is given by the Lexis formula:¹

$$\sigma_{p_k}^2 = \frac{PQ}{s_k} + \frac{s_k - 1}{ns_k} \sum_{t=1}^n (P_t - P)^2 \tag{1}$$

in which the P_t represent the respective probabilities that the fertile eggs produced by the n birds will hatch, P is the arithmetic mean of the P_t , and Q is equal to $1 - P$.

The values of the probabilities, P_t , are not known. However, as a first approximation to equation (1) we may write:

$$\sigma_{p_k}^2 = \frac{pq}{s_k} + \frac{s_k - 1}{ns_k} \sum_{t=1}^n (p_t - p)^2 \tag{2}$$

in which p is the arithmetic mean of the p_t and q is equal to $1 - p$.

The product, pq , can be accepted as a reasonably close approximation to the product, PQ , but the expression, $\sum_{t=1}^n (p_t - p)^2$, will, in general, be greater than the expression, $\sum_{t=1}^n (P_t - P)^2$. The reason for this is apparent when we consider that if each of these two expressions is divided by n , the former yields an estimate of the squared standard deviation of the p_t while the latter yields an estimate of the squared standard deviation of the P_t . The standard deviation of the p_t will, in general, be greater than that of the P_t because the p_t are more or less imperfect estimates of the P_t and are, therefore, subject to sampling errors from which the P_t are free.

We may write:

$$\frac{1}{n} \sum_{t=1}^n (P_t - P)^2 = \frac{1}{n} \sum_{t=1}^n (p_t - p)^2 - \sigma_c^2 \tag{3}$$

in which σ_c^2 is an appropriate correction as yet undefined.

Since the p_t would approach the P_t as statistical limits if each of the s_t were made extremely large, it follows that σ_c^2 must approach zero as each of the s_t approaches infinity. Furthermore, if $P_1 = P_2 = \dots = P_n = P$, we must have:

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n (p_t - p)^2 - \sigma_c^2 &= 0 \quad \text{or} \\ \sigma_c^2 &= \frac{1}{n} \sum_{t=1}^n (p_t - p)^2. \end{aligned} \tag{4}$$

¹ The formula as given in this paper is a modification of that given by Rietz, H. L. (1927) in his book, *Mathematical Statistics*, Open Court Publishing Co., Chicago, which was necessary in order to make it applicable to relative frequencies.

These conditions suggest that σ_c^2 be defined by the equation:

$$\sigma_c^2 = \frac{pq}{n} \sum_{t=1}^n \frac{1}{s_t}. \quad (5)$$

If σ_c^2 is so defined, it will obviously approach zero as each of the s_t approaches infinity. Furthermore, it has been shown by Yule² that if we have a series of n relative frequencies, such as the p_t under discussion, based on n samples of unequal size, and the probabilities of the occurrence and non-occurrence, respectively, of the particular event under consideration are constant from sample to sample, the squared standard deviation of those relative frequencies is given by a relation such as that used to define σ_c^2 in equation (5). Therefore, the second condition is also satisfied. σ_c^2 may be interpreted as representing that part of the squared standard deviation of the p_t which is due to the unreliability of the p_t as estimates of the P_t .

Therefore, it seems reasonable to write:

$$\frac{1}{n} \sum_{t=1}^n (P_t - P)^2 = \frac{1}{n} \sum_{t=1}^n (p_t - p)^2 - \frac{pq}{n} \sum_{t=1}^n \frac{1}{s_t}. \quad (6)$$

Combining equations (1) and (6), we obtain the following formula for calculating the squared standard error of p_k :

$$\sigma_{p_k}^2 = \frac{pq}{s_k} + \frac{s_k - 1}{ns_k} \left[\sum_{t=1}^n (p_t - p)^2 - pq \sum_{t=1}^n \frac{1}{s_t} \right]. \quad (7)$$

Since the weight of a measurement is inversely proportional to the square of the standard error of the measurement, we are now in a position to calculate a weighted mean, \bar{p} , of the p_t .

$$\bar{p} = \frac{\sum_{t=1}^n w_t p_t}{\sum_{t=1}^n w_t} \quad (8)$$

in which:

$$w_t = \frac{1}{\sigma_{p_t}^2}. \quad (9)$$

The squared standard error of \bar{p} is given by the familiar formula:

$$\sigma_{\bar{p}}^2 = \frac{\sum_{t=1}^n w_t (p_t - \bar{p})^2}{(n - 1) \sum_{t=1}^n w_t}. \quad (10)$$

² Yule, G. Udny, 1927. Introduction to the Theory of Statistics, Charles Griffin and Co., London.

It would seem that \bar{p} may be accepted as a good estimate of the average hatchability of the fertile eggs produced by the given lot of birds, and that equation (10) may be used to obtain a valid estimate of the reliability of \bar{p} .

However, the problem is not quite so simple. In the first place, there is usually a small amount of positive correlation between the number of fertile eggs produced by a bird and the hatchability of those eggs. Secondly, as pointed out earlier in this paper, the hatchability of fertile eggs is influenced to some extent by the male birds used to fertilize the eggs. The error involved in neglecting the correlation between hatchability and number of fertile eggs incubated does not seem to be of much importance in those practical problems which have come to the author's attention. The effects of differences among the male birds may be largely obviated in experimental work by frequently transferring male birds from lot to lot during the experimental period.

The best test of the suitability of a particular formula for calculating the standard error of an average is to compare the value of the standard error calculated by means of the formula with the corresponding value obtained by direct calculation from the distribution of a number of such averages obtained under essentially the same conditions. The accompanying table gives the standard error of the weighted average hatchability of fertile eggs calculated for each of four lots of birds by means of equation (10), together with the corresponding values obtained from the distribution of averages. The former are designated as the "predicted" values and the latter are designated as the "observed" values. In the calculation of the "observed" values, the various averages were assigned the same weights which were used in the calculation of the "predicted" values.

Comparison of "predicted" and "observed" standard errors of the weighted average hatchability of fertile eggs, calculated for each of four lots of birds

Lot	\bar{p}	Standard error of \bar{p}	
		"Predicted"	"Observed"
1	0.7684	0.0287	0.0327
2	0.7115	0.0533	0.0561
3	0.6834	0.0355	0.0379
4	0.7260	0.0615	0.0674

The data used in these calculations involved a total of 74 birds, approximately equally divided among the four lots, and a total of 2,901 fertile eggs which were produced and incubated during an experimental period of 48 weeks. The agreement between the "predicted" and "observed" standard errors of the weighted average hatchability for each lot of birds is excellent. However, the author's experience with biological data tends to make him doubt that such

close agreement will always be found when such data are subjected to the above treatment. The agreement in the present illustration could be less close without indicating that the method of calculating the "predicted" standard errors is unsound.

BUREAU OF ANIMAL INDUSTRY,
U. S. DEPARTMENT OF AGRICULTURE,
WASHINGTON, D. C.