NOTES ON STATISTICS

By Dr. Charles H. Blake

These notes will cover the main material for the statistics workshop at the 1968 EBBA annual meeting. My hope is that they will stimulate questions.

Suppose we have a collection (or array) of observations of a characteristic of a species. We call each observation a "variate". Any variate has a numerical value x. The number of variates of the same x is its frequency, f. The total number of variates is N. The mean or arithmetical average value of the array of variates is $\bar{X}$. (Read "bar-X".)

Our problem is to choose a proper mathematical model to represent the whole population of which the observed variates are a sample. This model is a "statistical distribution". From the sample we deduce estimates of various properties of the distribution.

The mean ($\bar{X}$) is the best representative value of the sample and estimates the corresponding value of the distribution. The standard deviation (s) measures the spread or dispersion of the values of the variates and estimates the dispersion of the distribution.

The Gaussian distribution (often called "normal"): variates may have any value, integral or fractional, positive or negative. In the typical

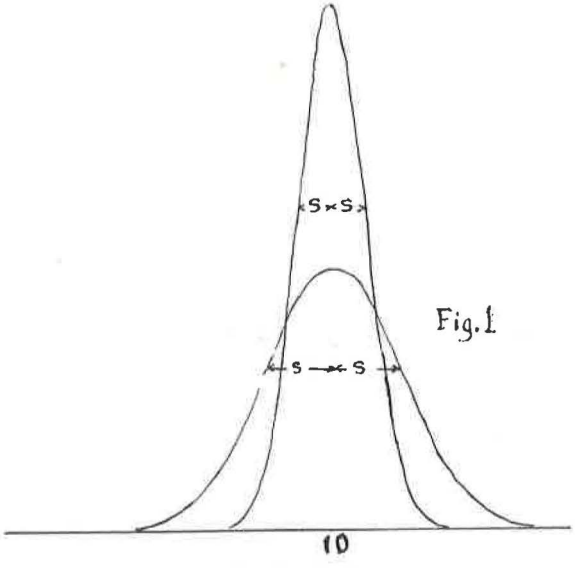

Fig. 1

Fig. 1. Gaussian distributions with same number of variates and same mean, $s = 1$, and $s = 2$.

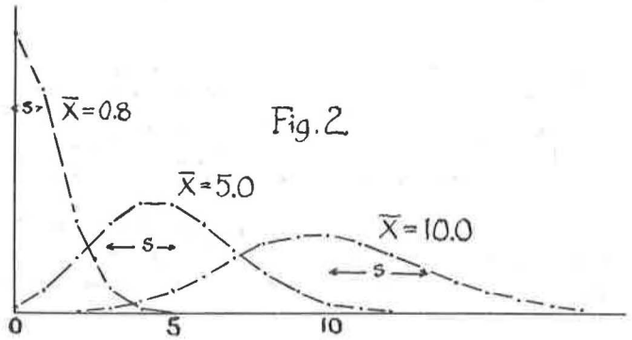

Fig. 2.

Fig. 2. Poisson distributions with means as shown, all with same number of variates.

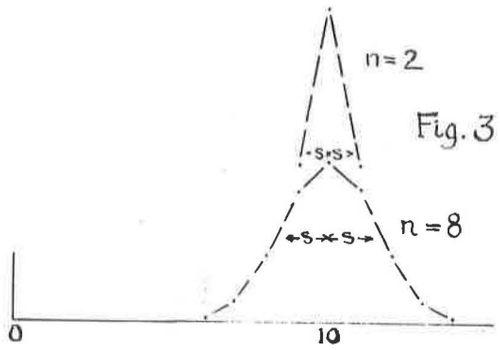

Fig. 3.

Fig. 3. Bernouilli distributions with same mean and number of variates, n = 2, and n = 8.

case the distribution is symmetrical and defined by two parameters, $\bar{X}$ and s. Then the skewness is 0 and the kurtosis 3.0. Example: wing lengths, Fig. 1.

The Poisson distribution: variates must be integral and zeo or positive. The distribution is asymmetrical and defined by one parameter, $\bar{X}$. Example: numbers of Cowbird eggs in host nests. Fig. 2.

The Bernouilli distribution: variates must be integral and usually are positive. The distribution is summetrical in the typical case. A convenient expression for this distribution is $(a + b)^n$. The parameters are a or p, the probability of occurrence of a; b or q, the probability of occurrence of b; and n. In all cases $p + q = 1$. In symmetrical cases $p = q$. The numbers of variates of any value are proportionate to the coefficients of the expanded polynomial. In all cases the number of terms of the expansion is n + 1. In symmetrical cases the sum of the coefficients is 2^n if we use a and b or 1 if we use p and q.

Numerical example (1) $a = b = 1$, $n = 3$ $(p = q = \frac{1}{2})$

$$\begin{array}{l} 1 + 1 \\ \underline{1 + 1} \\ 1 + 1 \\ \quad \underline{1 + 1} \\ 1 + 2 + 1 \\ \quad \underline{1 + 1} \\ \quad 1 + 2 + 1 \\ \quad \quad \underline{1 + 2 + 1} \\ 1 + 3 + 3 + 1 \end{array}$$

Sum of coefficients = $8 = 2^n$

If we have 96 observations then the actual numbers (frequencies) are: 12, 36, 36, 12

$s = \sqrt{\frac{1}{2} \times \frac{1}{2} \times 2} = \sqrt{\frac{1}{2}} = 0.71$

Numerical example (2) $a = 2$, $b = 1$, $n = 4$ $(p = 2/3, q = 1/3)$

$$\begin{array}{l} 2 + 1 \\ \underline{2 + 1} \\ 4 + 2 \\ \quad \underline{2 + 1} \\ 4 + 4 + 1 \\ \underline{4 + 4 + 1} \\ 16 + 16 + 4 \\ \quad 16 + 16 + 4 \\ \quad \quad \underline{4 + 4 + 1} \\ 16 + 32 + 24 + 8 + 1 \end{array}$$

Sum of coefficients = 81

If we have 100 variates, the frequencies are: 19.8, 39.5, 29.7, 9.9, 1.2

$s = \sqrt{2/3 \times 1/3 \times 4} = \sqrt{8/9} = \sqrt{.89} = .94$

This is the essential distribution in genetics but not common in other aspects of biology. When you do need it, you need it badly. The position on the X-axis is determined by the integers (successive) assigned to the variates and these same values are used in computing $\bar{X}$. Fig. 3.

We now need the operator Σ (sigma) which means "obtain the sum" and the quantity d, the deviation of any variate from the mean. For the Gaussian distribution, $s = \sqrt{\Sigma d^2/N}$ or the square root of the mean of the squares of the deviations. Croxton gives a more convenient method of calculating s. For the Poisson distribution $s = \sqrt{X}$. For the Bernouilli distribution $s = \sqrt{pqn}$.

Notes on the Figures: The Gaussian distributions, Figs. 1, 4 and 5 are in solid lines because the distribution is continuous. The Poisson

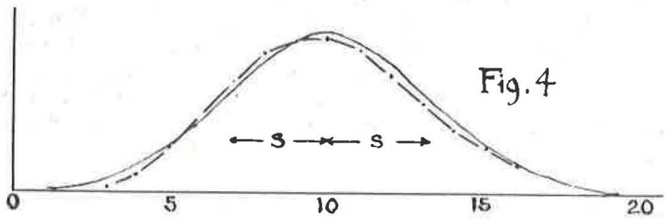

Fig. 4. See text, $\overline{X} = 10$, $s = 3.2$ for both curves.

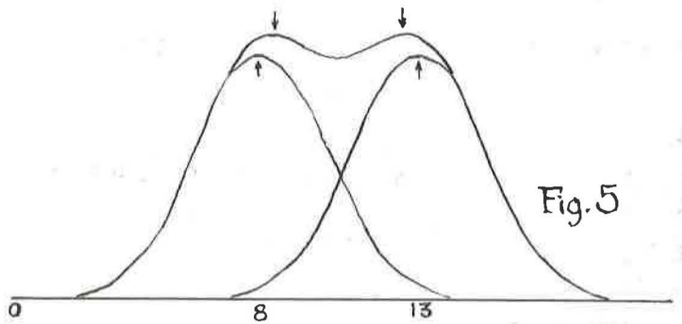

Fig. 5. See text, means of components 8 and 13, peaks of summed curve, approx. $8\frac{1}{2}$ and $12\frac{1}{2}$.

distributions, Figs. 2 and 4, and the Bernouilli, Fig. 3 are plotted by dots tied together with dashes. These are discontinuous distributions. Fig. 4 shows that, if the mean and standard deviation are the same and the mean is large enough, then the Gaussian and Poisson distributions are nearly the same.

Fig. 5 is to illustrate an often overlooked point. If two overlapping Gaussian distributions are added together the resulting distribution may have either one or two peaks depending on the distance between the means of the components. If there are two peaks, they will be separated by less than the distance between the means of the components. If the separation of the peaks is zero then the summed curve has only one broadened peak.

The term "sample size" refers only to the number of variates contained in the sample. A much misused statistic is the "standard error of the mean". This is the standard deviation of an array of means of samples, all of the same size, drawn from a population. Its estimate is $s_{\bar{x}} = s/\sqrt{N}$. (s divided by square root of N -Ed.) It is useful in deciding whether two samples of the same size were drawn from the same or different populations. It cannot be used when the sample sizes are different.

As far as we have carried the matter here, statistics provides procedures for obtaining useful information from arrays of observations. Statistics cannot develop information which is not in the observations. The accuracy of the information depends on a correct choice of statistical procedures. The arithmetic involved is sometimes tedious but seldom difficult.

Literature - grading in general from less to more difficult:

Croxton, F.E., 1959. Elementary Statistics. Dover Public.

Larsen, H.D., 1948. Rinehart Mathematical Tables. Rinehart & Co.

Defense Syst. Dept., G.E. Co., 1962. Tables of the individual and cumulative terms of the Poisson Distribution. Van Nostrand.

Moroney, M.J., 1959. Statistical methods in bilogy. Engl. Univ. Press.

Wilks, S.S., 1951. Elementary Statistical Analysis. Princeton U. Press.

Fisher, R.A., 1948. Statistical methods for research workers, ed. 10., Hafner, N.Y.
1949. The design of experiments, ed. 5. Hafner, N.Y.

Rao, C.R., 1952. Advanced statistical methods in biometric research. Wiley.

Box 613, Hillsborough, North Carolina 27278