THE DISPERSION OF OBSERVATIONS.

BY PROFESSOR JULIAN LOWELL COOLIDGE.

In the study of statistics bearing upon groups of objects, it is of fundamental importance to know whether, and to what extent, the different objects in the same group are comparable, and also to what extent one group is comparable with another. The first writer to study such questions seems to have been Lexis* and subsequent writers have developed his ideas of normal, supernormal and subnormal dispersion, or of Bernoulli, Lexis, and Poisson series.

In all articles dealing with these topics, which have come to the writer's attention, the data are either restricted to series of probabilities or frequency ratios, or else it is assumed that variations are distributed according to the Gaussian law of error. This restriction is unnecessary and unfortunate. The purpose of the present paper is to show how any series of sets of observations may be tested for the normality of their dispersion. The mathematical means employed are of the most elementary nature.

Suppose that we have n observed values $y_1, y_2, \cdots, y_n$. The expression

$$\sqrt{\frac{1}{n}\left[\sum_i\left(y_i - \sum_j \frac{y_j}{n}\right)^2\right]}$$

is called the *dispersion* or *standard deviation* of the set, and is of first importance in the theory of errors of observation, and of statistics. Let us find the mean value of its square. For simplicity, the mean value of y_i shall be called a_i while the mean value of its square is A_i. We shall write the averages,

$$(1/n)\sum y_i = y, \qquad (1/n)\sum a_i = a.$$

The square of the dispersion is

$$(1/n)\sum_i (y_i - y)^2 = (1/n)\sum_i \left\{\Big((y_i - a_i) - (y - a)\Big) + [a_i - a]\right\}^2.$$

In the large round parenthesis, if the averages y and a be replaced by their expanded values, we have the sum of a number of terms, each of which has the mean value 0, and as these terms are independent, the mean value of the product

* *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft*, Freiburg, 1877.

of two is also 0. The mean value of the square of the round bracket is the sum of the mean values of the squares of its individual terms. The mean value of the product of the round and square brackets is 0, and the mean value of the square of the square bracket is its ostensible value.

$$(y_i - a_i) - (y - a) = y_i - a_i - \frac{\sum(y_i - a_i)}{n}$$

$$= \frac{(n-1)}{n}(y_i - a_i) - \frac{(y_j - a_j) + (y_k - a_k) + \cdots (y_n - a_n)}{.\,n}.$$

The mean value of $(y_i - a_i)^2$ is $A_i - a_i^2$, and the coefficient will be $(n - 1/n)^2$ in one case, and $1/n^2$ in $(n-1)$ other cases. Summing, we reach the following theorem.

FUNDAMENTAL DISPERSION THEOREM. *If n independent quantities $y_1, y_2, \cdots, y_n$ be given, their mean values being $a_1, a_2, \cdots, a_n$, while the mean values of the squares are $A_1, A_2, \cdots, A_n$ respectively, and if $y = (1/n)\Sigma_i y_i$, $a = (1/n)\Sigma_i a_i$, then the mean value of the square of the dispersion is*

$$\frac{1}{n}\left[\frac{n-1}{n}\sum_i(A_i - a_i^2) + \sum_i(a_i - a)^2\right]\cdot$$

In practice it is convenient, first to replace $(n-1)/n$ by 1, and, secondly, to replace the mean value of the square of the dispersion by its observed value. We thus obtain the approximate equation

$$(1) \qquad \sum_i(y_i - y)^2 = \sum_i(A_i - a_i^2) + \sum_i(a_i - a)^2.$$

Let us show the practical application of this equation. Suppose that we have N sets, each of s observations,

$$\begin{array}{ll} x_{11}\ x_{12}\ \cdots\ x_{1s}; & \sum_j x_{1j} = x_1, \\ x_{21}\ x_{22}\ \cdots\ x_{2s}; & \sum_j x_{2j} = x_2, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ x_{N1}\ x_{N2}\ \cdots\ x_{N_s}; & \sum_j x_{Nj} = x_N. \end{array}$$

Let the mean value of x_{ij} be a_{ij}, while the mean value of its square is A_{ij}; and let $\Sigma_j a_{ij} = a_i$, $\Sigma_i x_i = Nx$, $\Sigma_i a_i = Na$. Then, by (1), we have

$$\sum_j\left(x_{ij} - \frac{x_i}{s}\right)^2 = \sum_j(A_{ij} - a_{ij}^2) + \sum_j\left(a_{ij} - \frac{a_i}{s}\right)^2\cdot$$

Summing again, we find

$$(2) \qquad \sum_{i,j}\left(x_{ij} - \frac{x_i}{s}\right)^2 = \sum_{i,j}(A_{ij} - a_{ij}^2) + \sum_{i,j}\left(a_{ij} - \frac{a_i}{s}\right)^2\cdot$$

Again, we may write $(x_i - a_i) = \Sigma_j(x_{ij} - a_{ij})$. The mean value of $(x_i - a_i)^2$ is $\Sigma_j(A_{ij} - a_{ij}^2)$. Hence applying (1) once more, we find

$$(3) \qquad \sum_i (x_i - x)^2 = \sum_{i,j} (A_{ij} - a_{ij}^2) + \sum_i (a_i - a)^2.$$

Eliminating $\Sigma_{ij}(A_{ij} - a_{ij}^2)$ between (2) and (3), we obtain

$$(4) \quad \sum_{i,j} \left[\left(x_{ij} - \frac{x_i}{s}\right)^2 - \left(a_{ij} - \frac{a_i}{s}\right)^2\right] = \sum_i \left[(x_i - x)^2 - (a_i - a)^2\right].$$

This is our fundamental equation. In practice it is usual to see whether a given series of sets of observations belong to one of three recognized types:

(A) Bernoulli series. Here all of the observations are supposed to be upon the same object, or at least the mean value does not vary from one object to another. We have

$$a_{ij} \equiv a_i, \quad a_i \equiv a, \quad \sum_{i,j} \left(x_{ij} - \frac{x_i}{s}\right)^2 = \sum_i (x_i - x)^2.$$

Such a series is said to have *normal dispersion.*

(B) Lexis series. All observations in one set are supposed to bear on the same object, but the object varies from set to set.

$$a_{ij} \equiv a_i, \quad a_i \not\equiv a, \quad \sum_{i,j} \left(x_{ij} - \frac{x_i}{s}\right)^2 < \sum_i (x_i - x)^2.$$

Such a series is said to have *supernormal dispersion.*

(C) Poisson series. The objects within a set are supposed to differ from one another, but all sets are supposed comparable.

$$a_{ij} \not\equiv a_i, \quad a_i \equiv a, \quad \sum_{i,j} \left(x_{ij} - \frac{x_i}{s}\right)^2 > \sum_i (x_i - x)^2.$$

Such a series is said to have *sub-normal dispersion.*

What we can do in practice is this. *We calculate the quantities $\Sigma_{i,j}\ (x_{ij} - x_i/s)^2$ and $\Sigma_i(x_i - x)^2$. If they be virtually equal, we are sure that the members of each set can not all be equal, unless the sets are all the same, and vice versa. If the first be less than the second, the various sets can not be the same. If the first be greater than the second, the individuals must differ within a set.*

It is well, in conclusion, to show how our formulas connect with the usual formulas for the dispersion of ratios. The problem is to see whether certain frequency ratios vary from case to case, or from set to set. If the generic letter for one of our probabilities be p_{ij}, and this represent the chance that

x_{ij} takes the value 1, while in the contrary case it takes the value 0, then a_i/s represents the average probability for the ith set.

$$\sum_j p_{ij} = sp_i = a_i \qquad \sum_i p_i = Np = a/s \qquad p + q = 1.$$

$$A_{ij} - a_{ij}^2 = \text{Mean value of } (x_{ij} - p_{ij})^2 = p_{ij} - p_{ij}^2.$$

From (3), we find

$$\sum_i (x_i - x)^2 = \sum_{i,j} (p_{ij} - p_{ij}^2) + s^2 \sum_i (p_i - p)^2.$$

Now

$$\sum_j (p_{ij} - p_i)^2 = \sum_j p_{ij}^2 - sp_i^2.$$

Hence

$$\sum_j p_{ij}^2 = sp_i^2 + \sum_j (p_{ij} - p_i)^2.$$

Similarly

$$\sum_i p_i^2 = Np^2 + \sum_i (p_i - p)^2,$$

$$\begin{aligned}\sum_i (x_i - x)^2 &= \sum_i [sp_i - sp_i^2 - \sum_j (p_{ij} - p_i)^2 + s^2(p_i - p)^2] \\ &= Nspq - \sum_{i,j}(p_{ij} - p_i)^2 + (s^2 - s)\sum_i (p_i - p)^2,\end{aligned}$$

$$(5)\quad \frac{1}{N}\sum_i (x_i - x)^2 = spq - \frac{1}{N}\sum_{i,j}(p_{ij} - p_i)^2 + \frac{s^2 - s}{N}\sum_i (p_i - p)^2.$$

To take up the particular cases:

(A) Bernoulli series. The probability is constant throughout.

$$p_{ij} \equiv p_i \equiv p, \quad \frac{1}{N}\sum_i (x_i - x)^2 = spq.$$

(B) Lexis series. The probability is constant throughout a set, but varies from set to set.

$$p_{ij} \equiv p_i, \quad p_i \not\equiv p. \quad \frac{1}{N}\sum_i (x_i - x)^2 = spq + \frac{s^2 - s}{N}\sum_i (p_i - p)^2.$$

(C) Poisson series. The probability varies within the set, but all sets are comparable.

$$p_{ij} \not\equiv p_i, \quad p_i \equiv p. \quad \frac{1}{N}\sum_i (x_i - x)^2 = spq - \sum_j (p_{ij} - p_i)^2.$$

These are the standard formulas.*

Harvard University,
April, 1921.

* Conf. Fisher, *Mathematical Theory of Probabilities*, New York, 1915, pp. 120 ff.