Separating Frequency Distributions into Two Normal Components

Arnold Court

Research and Development Branch,
Office of The Quartermaster General, Washington, D. C.

Separation of any frequency distribution into two component normal distributions may be effected readily if either of two assumptions can be made with validity: (1) that the means M_1 and M_2 of the components are known; (2) that the standard deviations of the components are equal.

The only other requirements for the separation are knowledge of the total frequency (or area) N of the given distribution and its mean M, standard deviation σ, third moment ν_3, and, for the second case only, fourth moment ν_4. The method for such separation, outlined by Charlier (*1*) more than 40 years ago but little noticed, is based on Pearson's (*5*) general method for finding two normal components in any distribution, which assumes nothing about them except their existence, and requires solution of a complete ninth degree equation involving the first five moments of the given distribution.

Pearson applied his method not only to markedly skewed distributions, in which the presence of two components is indicated strongly, but to some which are quite symmetrical (although not normal), to find components with identical means but differing standard deviations. His general method applies even when one of the components is negative, i.e., when the given distribution is the difference between two normal ones. This method was used by Crum (*2*) and Pollard (*3*), without reference to De Helguero's (*3*) slight modifications, to Pearson's (*7*) own refinement, or to the simple solutions developed by Charlier for two special cases of great practical significance.

Assumed Means. In obviously bimodal distributions, and many unimodal ones with pronounced "humps" or "shelves," means M_1 and M_2 for two supposed components may be apparent. Their departures from the mean M of the given distribution,

$$M - M_1 = m_1 \text{ and } M_2 - M = m_2$$

were used by Charlier (*1*) to find the variances of the two components (notation simplified):

$$\sigma_1^2 = \sigma^2 - 2m_1m_2/3 - (m_1^2/3 + \nu_3/3m_2)$$
$$\sigma_2^2 = \sigma^2 - 2m_1m_2/3 - (m_2^2/3 - \nu_3/3m_1).$$

The total areas or frequencies of each component depend only on the assumed means:

$$N_1 = Nm_2/(m_1 + m_2) \qquad N_2 = Nm_1/(m_1 + m_2).$$

Finally, from a table of normal frequency curve ordinates, $\phi(t)$, the ordinate at any distance (in t units) from the mean may be found, since

$$y_1 = (N_1/\sigma_1)\,\phi(t).$$

The larger component always corresponds to the smaller departure from the mean, which in turn is m_1 if ν_3 is positive, m_2 if negative. Should impossible means be assumed for the two components, σ_1^2 or σ_2^2 will be negative, indicating no real solution.

However, the method of assumed means does not give a unique solution: usually trial of several pairs of means is required to find one set yielding two components which, added together, closely approximate the given distribution. The best pair of means generally has maximum ordinates agreeing well with the observed values, due regard being given to the contribution each component makes to the other's peak. Such agreement can be made

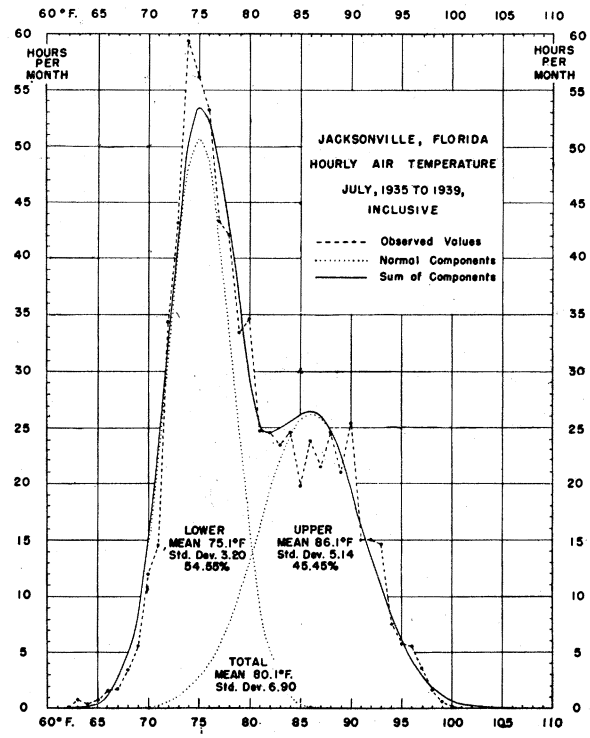

FIG. 1.

as close as desired by assuming values of the maximum ordinates y_1 and y_2 in addition to the means M_1 and M_2. Then

$$\sigma_1 = N_1/\sqrt{2\pi}\,y_1 \quad \text{and} \quad \sigma_2 = N_2/\sqrt{2\pi}\,y_2.$$

Example. Charlier's method of assumed means (without assuming values for the maximum ordinates) has been applied (Fig. 1) to 3720 hourly temperature readings during July at Jacksonville, Florida; significance of the two components found by this method will be discussed elsewhere. In the data, even temperatures are disproportionately more frequent than odd ones, because observers, required to estimate readings to tenths of a degree, tended to values ending in 0.0 or 0.5; the latter, by the classic rule for disposal of decimals, then were rounded to the nearest even degree. If, instead, all 0.5 values were rounded to the *next higher* units, frequency distributions would be smoother and means too high by only 0.05°.

For these Jacksonville data, several pairs of assumed means were tried and discarded before two finally were found yielding components whose sum (Fig. 1) seems a

reasonable approximation to the original curve. A still better fit might have been obtained from means differing from the given mean by tenths of a degree, instead of by the half-degree intervals fixed upon to simplify computations.

Equal standard deviations. Assuming the two presumed components to have equal standard deviations, instead of assuming values for their means, led Charlier (*1*) to a cubic equation involving the difference between the variances of the given distribution and the assumed components:

$$z^3 + \tfrac{1}{2}(\nu_4 - 3\sigma^4)z + \tfrac{1}{2}\nu_3^2 = 0,$$

where $z = \sigma_1^2 - \sigma^2$. The discriminant of this cubic,

$$C^2 = (\sigma^{12}/216)\ (13.5\alpha_3^4 + E^3),$$

where $\alpha_3 = \nu_3/\sigma^3$ is the skewness and $E = (\nu_4/\sigma^4) - 3$ the excess, almost always is positive, indicating only one real root:

$$z = 0.4082\ \sigma^2\ (\sqrt[3]{-3.6742\alpha_3^2 + \gamma}\ -\ \sqrt[3]{3.6742\alpha_3^2 + \gamma}),$$

where $\gamma = \sqrt{13.5\alpha_3^4 + E^3}$.

Except for almost symmetrical and very flat-topped distributions, γ is positive, so that z will be negative, and

$$\sigma_1^2 < \sigma^2.$$

But if $-z > \sigma^2$, then σ_1^2 is negative, and there is no actual solution, indicating the assumption of equal standard deviations to be unwarranted; for Jacksonville July temperatures, variances assumed to be equal are -5.07.

If the assumption is justified, and σ_1 is real, the means are:

$$M_1 = M - m_1 = M - (\nu_3/6) - \sqrt{(\tfrac{1}{4}\nu_3)^2 - z}$$

$$M_2 = M + m_2 = M - (\nu_3/6) + \sqrt{(\tfrac{1}{4}\nu_3)^2 - z}.$$

The areas N_1 and N_2 of the two components are found as before.

An asymmetrical curve which is the sum of two normal curves "affords a good fit both to distributions which possess two distinct modes, and to skewed distributions with one mode" (*8*). That components have not been found generally for such distributions may be due to ignorance of Charlier's facile methods; this ignorance, in turn, may stem from Pearson's insistence (*6*), replying to Edgeworth's criticism (*4*), that his "process is not so laborious that it need be discarded for rough methods of approximation based upon dropping the fundamental nonic and guessing suitable solutions."

Of Charlier's methods, the first, that of assumed means, is far simpler than the second, which involves the fourth moment and a cubic equation. However, Charlier concentrated on the second, the "abridged method for dissecting frequency curves," because the cubic equation involved is actually one step in the general solution, "hence it is no loss of time to begin with this approximate method."

Charlier declared that the assumption that the standard deviations of the two components are equal "is of a more general character" than the assumed knowledge of the means of the two components. "Especially in biology it is a fairly probable supposition that two types found together in nature often possess *nearly* equal standard deviations," but "this abridged method is applicable only when there are a priori reasons for the assumption that the two components have nearly equal standard deviations." In many cases no such reasons exist, and then it is safer to assume certain values for the means of the components, especially when approximate means may be determined by inspection. In the example given, the approximate values of the means were obvious from the graph, and trial of a few pairs of values yielded one which gives a good fit, with markedly different standard deviations; assumption of equal deviations gave no solution.

Even closer agreement with the original distribution can be obtained if values may be assumed for the maximum ordinates as well as the means of the two components. In effect, this short cut replaces the standard deviation and skewness of the original distribution by a subjective evaluation which may be more effective for some distributions, but is not of as general applicability in finding two normal components.

References

1. Charlier, C. V. L. *Lunds Universitets Arsskrift,* ny följd, afdelningen 2, 1906, **1**, No. 5.
2. Crum, W. L. *J. Amer. statist. Ass.,* 1923, **18**, 607.
3. De Helguero, Fernando. *Biometrika,* 1905, **4**, 230.
4. Edgeworth, F. Y. *J. roy. statist. Soc.,* 1899, **62**, 125.
5. Pearson, Karl. *Philos. Trans. roy. Soc. London,* 1894, **185**, Series A, 71.
6. ———. *Philos. Mag.,* 1901, **1**, 110.
7. ———. *Biometrika,* 1915, **10**, 479.
8. Pollard, Harry S. *Ann. math. Statist.,* 1934, **5**, 227.

Combination of Tissues from Different Species in Flask Cultures[1]

Clifford Grobstein and J. S. Youngner[2]

National Cancer Institute, Bethesda, Maryland

Combining tissues from different species apparently has been performed only infrequently in tissue culture. Roffo (*5*) grew together chicken and rat tissues, both normal and neoplastic, without evidence of antagonism. In studies of the mode of transmission of the excitation involved in cardiac muscle contraction, several investigators (for reference see Leone [*3*]) combined embryonic heart fragments of the chick with similar fragments from other avian and mammalian species. The establishment of synchrony of beat was reported, and no mention was made of any incompatibility reactions.

Harris (*2*), in an attempt to determine whether direct incompatibility exists between tissues of different mammalian species in culture, paired heart, spleen, and kidney fragments from newborn mice and rats in roller tubes. Harris found that ". . . rat and mouse tissue cells are physiologically compatible *in vitro.*" In a related study, though not involving tissues from different species, Medawar (*4*) recently made combined fluid cultures of skin from two adult rabbits, between which skin grafts had failed to take, and found no evidence of incom-

[1] With the technical assistance of Clara Lee and Edward J. Soban.

[2] Present address: Department of Bacteriology, University of Pittsburgh Medical School.