DETERMINATION OF MATHEMATICAL STATISTICAL CHARACTERISTICS BY GRAPHICAL METHODS

by

J. Borján

Department of Building Materials, Technical University, Budapest

(Received February 15, 1973)
Presented by Prof. Dr. J. Talabér

1. Introduction

Besides calculation methods for determining mathematical statistical characteristics, graphical methods are also of interest because of their simple means required, their descriptiveness and exceptional rapidity.

This paper deals with the graphical determination of the parameters of one and two-parameter empirical distributions approaching normal (Gaussian) distribution, the graphical estimation of the parameters of regression functions with two or three variables and the criticism of applying the coefficient of correlation.

2. Graphical estimation of the parameters of Gaussian distribution

2.1 One-parameter Gaussian distribution

The so-called Gaussian paper [1], [2] is well known where the normal distribution function can be represented by a straight. There exists a co-ordinate system [3] where the frequency function of the normal distribution i.e. a "bell curve" will be straightened. For this the height of the frequency function must be reduced to unity.

By reduced frequency histogram a frequency diagram is meant, obtained by dividing the frequencies n_i by the maximum frequency $n_{\max}$

$$\varphi_i = \frac{n_i}{n_{\max}} 100\,(\%).$$

This linearizing network may be constructed [4] by making the straights representing the Gaussian distribution to compose a network of equilateral triangles. One vertex of the triangle is at the value of maximum probability. This network is to the scale of the Gaussian paper and united with it, and then the same two straights cut out the statistical characteristics readable off the frequency network and the distribution network.

The construction procedure of the network and the network itself are shown in Figs 1 and 2, respectively.

The reduced frequencies are plotted on scale φ_i. Equalizing straights intersect at the mean value. The same equalizing straights cut out the standard deviation from the straight σ. The scale p_i contains the values of the distribution function. For example: the probability of values less than 62.8 μsec to occur is 5%.

In general, the most frequent value (100%) needs not to be forced to coincide with the point of intersection of the straights imposed upon. Either an estimation is made for $n_{\max}$ or the point of intersection of the equalizing straights is shifted posteriorly to 100%.

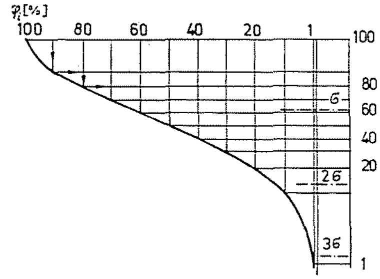

Fig. 1. Construction of the linearizing network of the frequency curve

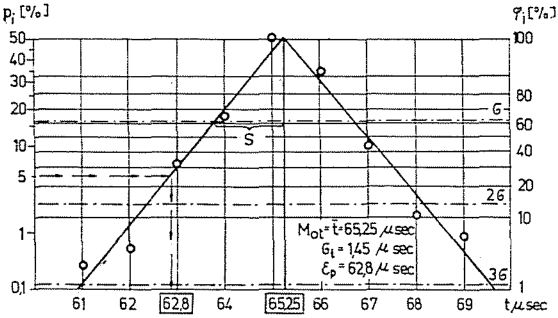

Fig. 2. United linearizing network of the frequency and the distribution function

If the value of $n_{\max}$ is previously fixed [5] the empirical frequencies may be plotted immediately on the network (Fig. 3).

(For each ξ value found, a graduation is passed forwards until reaching 100%.)

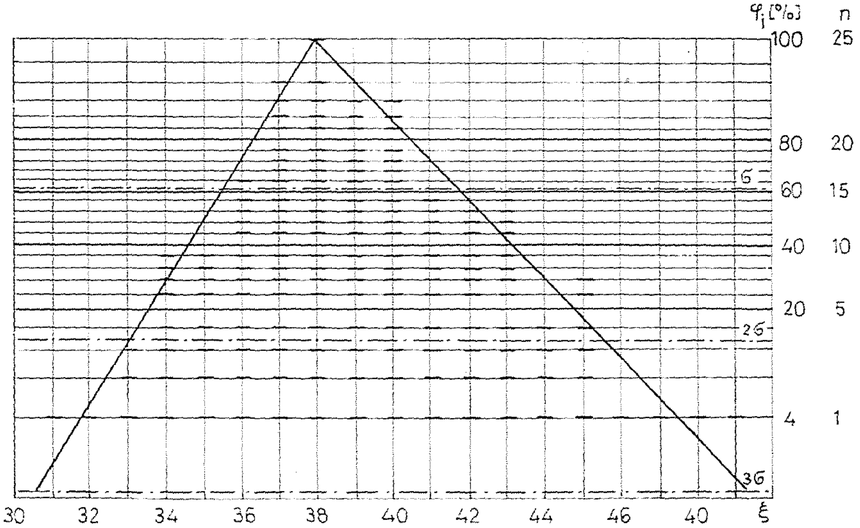

Fig. 3. Evaluation simultaneous to measurement

2.2 *Asymmetrical distribution*

Approaching asymmetrical empirical distributions with symmetrical Gaussian distributions may be erroneous. Therefore it is suggested [4] to approach the reduced empirical frequencies below and above the most probable value (modus) of normal distribution in the linearizing network using the "Gaussian half-straights" which cut out different standard deviations (Fig. 4).

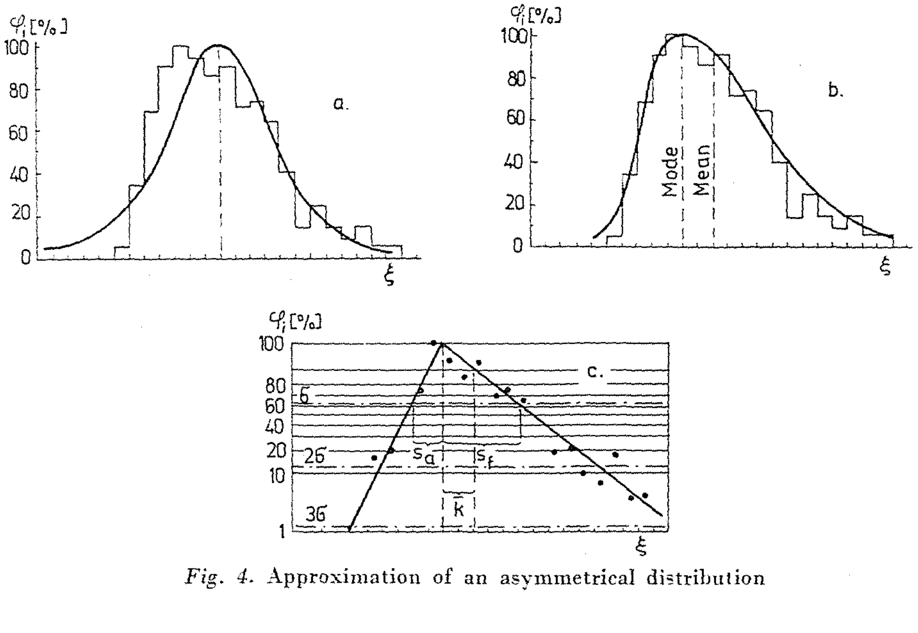

Fig. 4. Approximation of an asymmetrical distribution

This is allowed since reduced frequency functions are of unit height. The function will be defined by its modus (or mean value) as well as by the values of standard deviation below and above the modus s_a and s_f, respectively.

The mean value will be determined by calculation, the difference of mean value and modus being

$$\overline{K} = 0.7975\,(s_f - s_a)\,.$$

The standard deviation referred to the mean value is

$$s_0 + \sqrt{\frac{s_a^2 + s_f^2}{2}}\,.$$

The corrected value of probabilities:

$$p_i' = 2\,\frac{s_a}{s_a + s_f}\,p_i$$

where p_i means the probability level read off in case of symmetrical distribution. This function often fits better to an empirical frequency function than does the function of symmetrical Gaussian distribution.

2.3 *The modified Gaussian distribution*

Approximation by a Gaussian function can be verified by the central limit theorem. Nevertheless the Gaussian distribution has in many cases the disadvantage that the random variable generally cannot assume distribution values from $-\infty$ to $+\infty$. This presents a problem especially in estimating probability levels. For this reason several distribution types are constructed with zero loci not in the infinity. The zero loci of such distributions are included among the parameters of distribution function to be determined numerically.

To make use of the advantages of the graphical method for determining the parameters of a normal distribution, assumption of a distribution derived from the normal one with pre-estimable zero loci is suggested.

Since in general, the mathematical model describing the empirical function cannot be determined theoretically, the central limit theorem justifies mostly to assume normal or nearly normal distribution. In some cases the value of the zero locus of a function may be determined by theoretical considerations such as the minimum strength of an infinitesimal particle in statistical theories of failure.

The function proposed for the construction of the linearizing network is derived from the normal distribution with expected value of zero

$$f(\xi)=\begin{cases}\alpha\left[\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\xi}{\sigma}\right)^2}-\dfrac{\left|\dfrac{\xi}{\sigma}\right|}{\dfrac{a}{\sigma}}\cdot\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{a}{\sigma}\right)^2}\right]; & \text{for } \left|\dfrac{\xi}{\sigma}\right|\leq a\\[2em] 0 & \text{for } \left|\dfrac{\xi}{\sigma}\right|\geqslant a\end{cases}$$

where a/σ shows the multiple of the standard deviation of normal distribution the zero locus of the function lies at.

This function is presented in Fig. 5a for $a/\sigma = 2.545$ and 3.535 where σ is the standard deviation of the function. The second value is hardly different from the normal distribution.

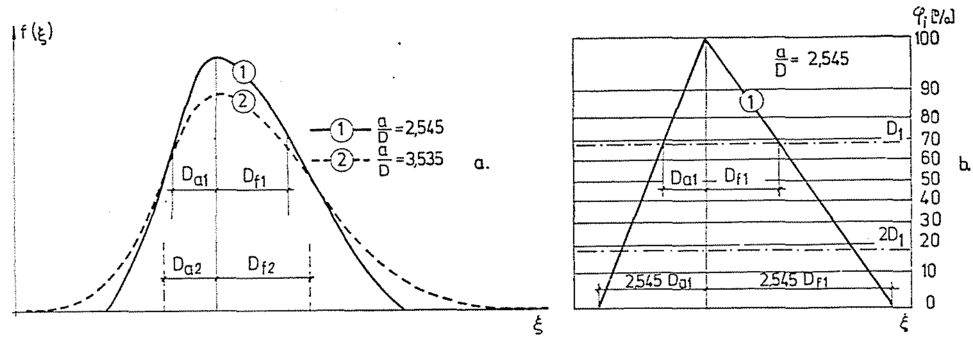

Fig. 5. Modified Gaussian distribution

For $a/\sigma = 2.545$ also the linearizing network has been constructed (Fig. 5b).

After plotting the reduced frequencies in network constructed for different a/σ values, the most suitable value may be selected by testing the fitting.

The modified Gaussian distribution fitting the empirical distribution is defined by its modus, mean value, standard deviation and by distance between zero locus and modus (distance in terms of standard deviation units).

2.4 *Two-parameter Gaussian distribution*

In case of two-parameter Gaussian distributions the fact is made use of that any vertical section of the frequency surface is limited by the frequency function of a one-parameter Gaussian distribution [1].

The construction (Fig. 6) consists in plotting for constant values of one variable (ξ_1) and constructing the reduced frequency values of the other independent variable ξ_2. All reduced frequency functions being of unit height, the mean values of the reduced frequencies can be plotted along the first variable, and connected with straights in the linearizing network. Repeating this in the other direction results in equalized straights corresponding to the frequency surface of the two-parameter distribution.

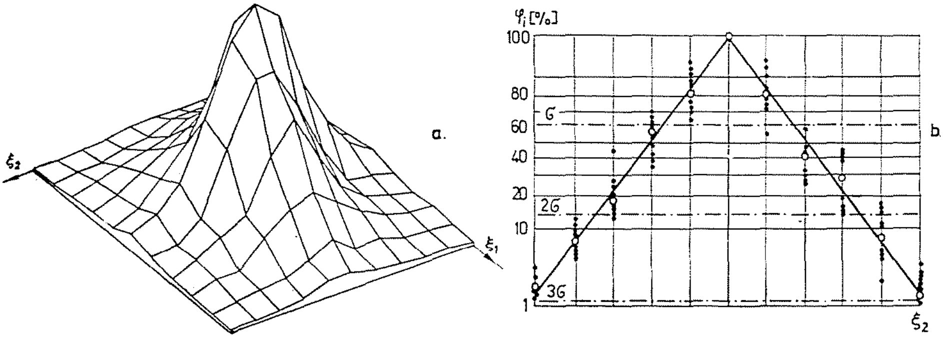

Fig. 6. Analysis of a two-parameter distribution

3. Graphic regression analysis

3.1 Finding relationship between the variables

At first — as in the numerical regression analysis — the kind of approximative function to be applied has to be decided. The law is either previously known or the function of appropriate form will be chosen by consideration, after having plotted the correlated test results.

Consider the case in [7] where the regression function is a straight (Fig. 7):

$$y = a \cdot x.$$

The slope of the straight line is considered as a random variable and denoted by ξ_{ai}.

The set of straights $y = \xi_{ai} \cdot x$, with slopes corresponding to random variable values ξ_{ai} is considered as class limits. Processing the frequencies of points between the class limits according to item 2.1, a_ξ, the most probable value of ξ_{ai}, will be obtained.

The equation of the regression function is

$$y = a_\xi \cdot x.$$

This approximation is correct if the standard deviation of the dependent variable varies linearily with x.

The residual standard deviation suitable to characterize the deviations from the function is given automatically.

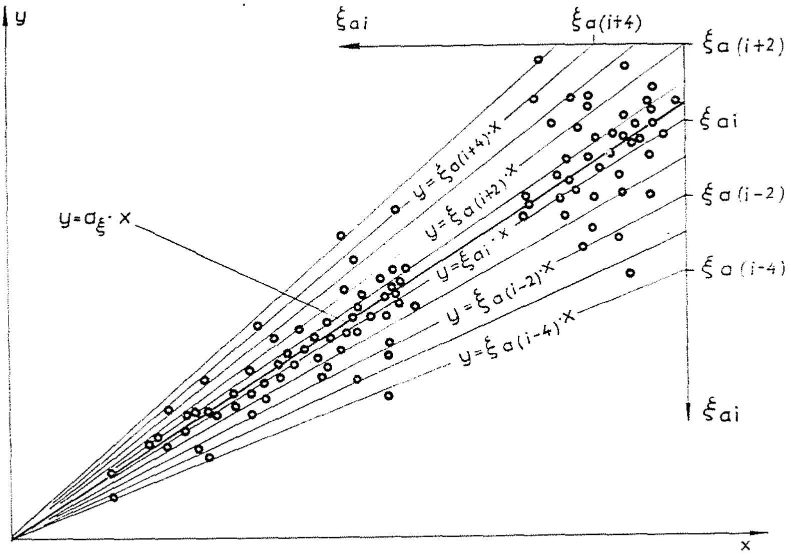

Fig. 7. Regression analysis with two variables

3.2 Regression analysis with three variables

Denote the three examined variables by α, β and γ (Fig. 8). Assume the rectangular co-ordinate systems $\alpha - \beta$ and $\gamma - \beta$ where test results will be plotted. The purpose is to construct in the $\alpha - \beta$ system the regression curve set expressing the relationship of the three variables so as to obtain the $\alpha - \beta$ curves of relation corresponding to the determined constant values of γ (Fig. 8c). The method is described in [8] with all particulars.

In the $\gamma - \beta$ system vertical straights correspond to regression curves γ_i (of any shape). Dividing the field of interpretation of γ into classes the points belonging to γ_i values can be found (Fig. 8b). Finding these points in the $\alpha - \beta$ system each regression curve γ_i can be traced, with shapes (tangents) and positions still susceptible to random variations. Along the curve plotted from the test results as two variables the position of the points of intersection and the slope of the tangents is equalized graphically (Fig. 8d) to deliver a set of regression curves taking also the three variables into account. It must be noted that the shape and tendency of the curves may rather deviate from the two-variable regression function plotted from the test results.

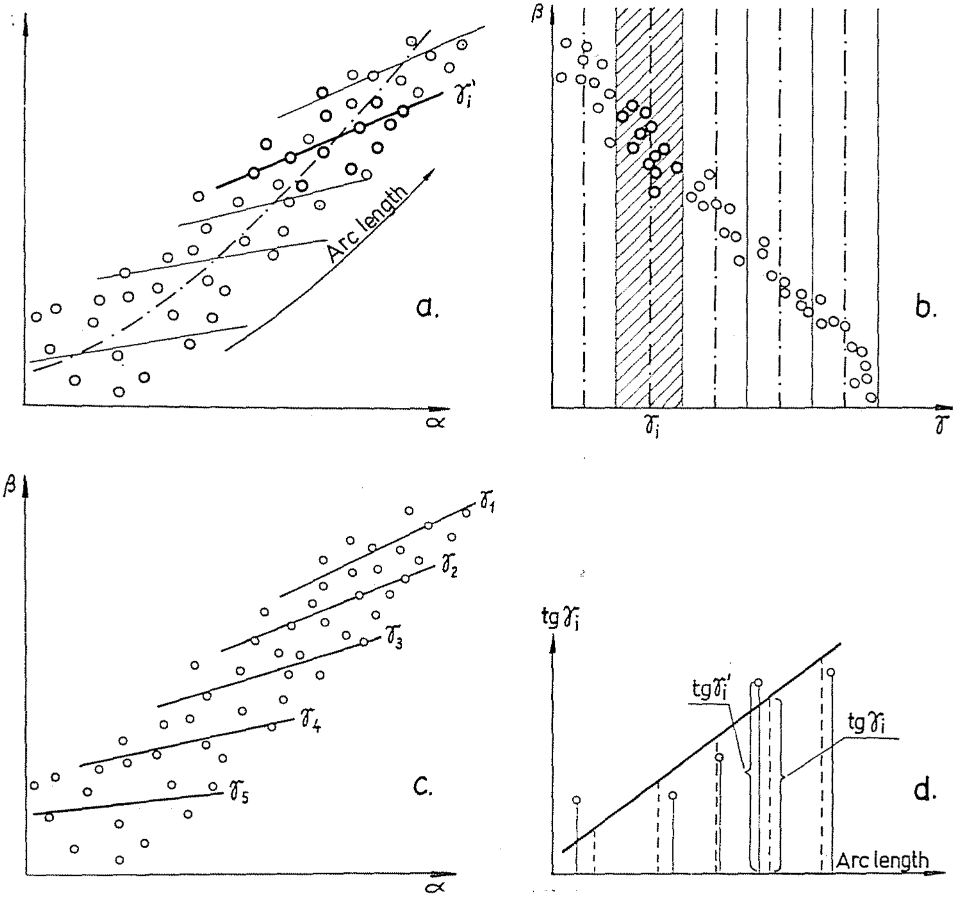

Fig. 8. Regression analysis with three variables

4. Examination of the correlation coefficient

The coefficient of correlation is expressed by

$$r_{xy} = \frac{s_{xy}}{s_x + s_y} .$$

The value of covariance s_{xy} in terms of standard deviation according to the variables x and y characterizes the closeness of the relation. (It is only valid for linear regression.)

The correlation is good for $r \cong 1$ while for $r \to 0$, the correlation is poor.

The value of s_{xy} will be high with a sign either + or — when most test results are grouped diagonally in quarters limited by straights drawn across their gravity lines with respect to x and y. The value will be low, near zero if the four fields are uniformly filled out by the test results.

$s_x \cdot s_y$ provides for non-dimensionality.

In Fig. 9 the correlation coefficient r_1 is seen to hardly differ from 1.

If only a part of the x value set (Fig. 9b) is available then — according to the above — the coefficient of correlation will be lower, in spite of a constant range of scatter (deviation from the function — dotted line) that is, of an unchanged probability of deviations from the regression function.

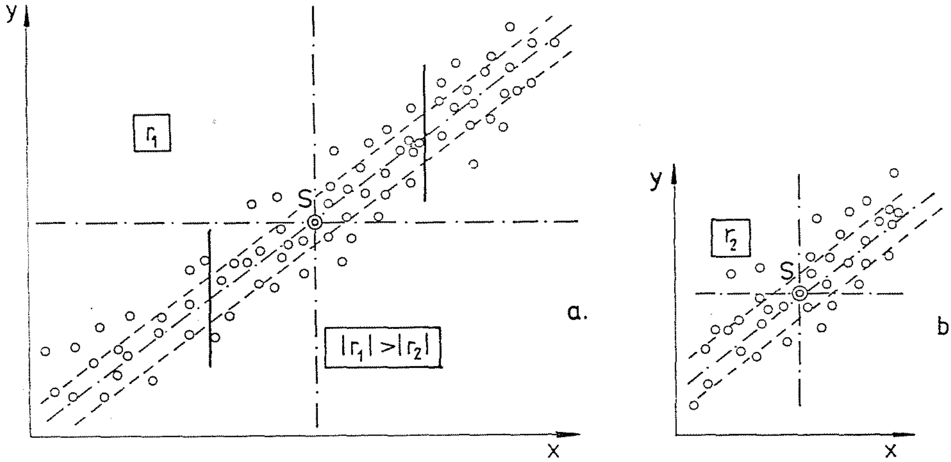

Fig. 9. Examination of the coefficient of correlation

The use of the coefficient of correlation is therefore restricted to the comparison of reliability of regression functions obtained from test results with identical value sets.

The residual standard deviation of the function is independent both of the value set and of the shape of the curve.

Summary

Normal distribution parameters (mean, modus, standard deviation, probability levels) may be graphically determined by plotting the reduced frequency histogram and the distribution histogram in a unified linearizing network.

Asymmetrical distributions may be approached by Gaussian half-curves having different standard deviation below and above the modus.

As mathematical model for near to normal empirical distributions a distribution derivable by simple operations from the normal one, satisfying the condition that the zero values of the function are not in the infinity, is suggested.

These methods have the advantage of simplifying the determination of the frequency surfaces of empirical distributions with two parameters, even of those with two-way different asymmetries.

The constants of two-variable regression functions are considered temporarily as random variables directly obtained graphically from the test results. The most probable values of the variables are the parameters of regression function. The method is also valid for a varying standard deviation of the dependent variable along the regression function.

If three variables are taken simultaneously into consideration — as against other methods for similar purposes — the section curves of the regression surface are obtained by classifying with respect to the third variable and successively, graphically equalizing the parameters of regression functions.

The correlation coefficient is suitable only for comparing the correlation degree of populations with identical value sets.

References

1. Rényi, A.: Probability Calculus.* Budapest, 1954. Tankönyvkiadó.
2. Felix, M.—Blaha, K.: Mathematical Statistics in the Chemical Industry.* Budapest, Műszaki Könyvkiadó, 1964.
3. Leinweber: Length Measurement.* A pocket book. Műszaki Könyvkiadó, Budapest, 1960.
4. Borján, J.: Rapid Graphical Determination of Mathematical Statistical Characteristics.* Mélyép. Szle. Vol. **18.** No. 7. 1968. pp. 289—293.
5. Borján, J.: Mathematical Statistics for Evaluating Non-Destructive Tests of Concrete.* Mélyép. Szle. Vol. **18.** No. 7, 1968. pp. 294—297.
6. Borján, J.—Szittner, Gy.: Approximation of Empirical Distributions by a Modified Gaussian Distribution.* Manuscript.
7. Borján, J.: Estimation of the Parameters of Regression Functions by a Graphical Method.* Mélyép. Szle. Vol. **21,** No. 1., 1971, pp. 32—38.
8. Borján, J.: Multiple Regression Analysis by a Graphical Method.* Manuscript.

Sen. Ass. József Borján, 1111 Budapest, Műegyetem rkp. 3. Hungary

* In Hungarian.