METRON

RIVISTA INTERNAZIONALE DI STATISTICA — REVUE INTERNATIONALE DE STATISTIQUE
INTERNATIONAL REVIEW OF STATISTICS — INTERNATIONALE STATISTISCHE ZEITSCHRIFT

DIRETTORE PROPRIETARIO — DIRECTEUR ET PROPRIÉTAIRE
EDITOR AND PROPRIETOR — HERAUSGEBER UND EIGENTHÜMER

Prof. Dott. Corrado Gini, *Direttore dell'Istituto di Statistica della R. Università di Roma.*

COMITATO DIRETTIVO — COMITÉ DE DIRECTION
EDITORIAL COMMITTEE — DIREKTIÔN-KOMITEE

Prof. F. Bernstein, *früherer Direktor des Instituts für mathematische Statistik der Universität, Göttingen* (Deutschland), *jetzt an der New York University* (U. S. A).
Prof. A. E. Bunge, *ya director gen. de Estadística de la Nación, Buenos Aires* (Argentina).
Prof. F. P. Cantelli, *professore di Matematica Attuariale nella R. Università di Roma* (Italia).
Prof. A. Flores de Lemus, *jefe de Estadística del Min. de Hacienda, Madrid* (España).
Prof. M. Greenwood, *professor of Epidemiology and Vital Statistics in the University of London* (England).
Dott. G. Jahn, *directeur du Bureau Central de Statistique de Norvège, Oslo* (Norvège).
Prof. A. Julin, *secrétaire général honoraire du Ministère de l'Industrie, du Travail et de la Prévoyance sociale, Bruxelles* (Belgique).
Prof. H. W. Methorst, *directeur de l'Office permanent de l'Institut International de Statistique et du Bureau central de Statistique, La Haye* (Pays-Bas).
Prof. W. F. Ogburn, *professor of Sociology in the University of Chicago* (U. S. A.).
Prof. R. Pearl, *director of the Department of Biology of the School of Hygiene and Public Health, Baltimore* (U. S. A.).

AMMINISTRATORE — ADMINISTRATEUR — MANAGER — VERWALTER
Prof. Silvio Orlandi, *Istituto di Statistica della R. Università di Roma.*

SEGRETARI DI REDAZIONE — SECRÉTAIRES DE RÉDACTION
EDITORIAL SECRETARIES — REDACTIONSSECRETAERE
Prof. Vittorio Castellano — Prof. Mario Saibante

ISTITUTO GENERALE DI STATISTICA — BIBLIOTECA — ROMA

Vol. XIII - N. 1 **28-II-1937.**

SOMMARIO — SOMMAIRE — CONTENTS — INHALT

C. F. Roos. *A general invariant criterion of fit for lines and planes where all variates are subject to error* . . . . . Pag. 3
H. E. Jones. *Some Geometrical Considerations in the General Theory of Fitting Lines and Planes* . . . . . » 21
V. Castellano. *Sugli indici relativi di variabilità e sulla concentrazione dei caratteri con segno* . . . . . » 31
P. Ottestad. *The exponential frequency function and frequency distributions* . » 51
D. E. Scates and L. M. Van Nortwick. *The influence of restrictive routes upon the center of minimum aggregate travel* . . . . . » 77
Federici. *Ulteriori ricerche sugli studenti dell'Università di Roma* . . . . . » 93

ROMA
AMMINISTRAZIONE DEL «METRON»
R. UNIVERSITÀ — ISTITUTO DI STATISTICA

II 123 + 15

ARTICOLI GIUNTI ALLA RIVISTA CHE VERRANNO PUBBLICATI NEI PROSSIMI NUMERI.

(*Secondo l'ordine d'arrivo*).

ARTIKEL DIE AN DIE ZEITSCHRIFT ANGELANGT SIND UND WELCHE IN DEN NACHFOLGENDEN NUMMERN ERSCHEINEN WERDEN.

(*Nach der Reihenfolge des Eingangs*).

ARTICLES REÇUS PAR LA REVUE ET À PARAÎTRE PROCHAINEMENT.

(*D'après la date de reception*).

ARTICLES RECEIVED BY THE REVIEW WHICH WILL BE PUBLISHED IN FUTURE ISSUES.

(*According to date of receipt*)

C. Bruen — *Methods for the Combination of Observations: Modal Point or Most Lesser-Deviations, Median Loci or Least Deviations, Mean Loci or Least Squares, and Mid-Point of Least Range of Least Greatest-Deviation.*

A. Ollivier — *On the anaiysis of variation in general death rates.*

Gli Autori degli articoli inviati per la pubblicazione nella Rivista, rinunciano in favore della medesima alla proprietà letteraria degli articoli stessi, qualora vengano pubblicati.

The Authors of papers sent for publication in the Review are supposed to give up their copyright in favour of the Review if the papers are published.

Les Auteurs des articles envoyés à la Revue, pour y être publiés, renoncent, en faveur de celle-ci, à la propriéte littéraire de leurs articles, s'ils sont acceptés.

Die Verfasser der zur Veröffentlichung in der Zeitschrift zugesandten Aufsätze, werden, falls selbige veröffentlicht werden, auf ihre Verfasserrechte zu Gunsten der Zeitschrift verzichten mussen.

C. F. ROOS *

A general invariant criterion of fit for lines and planes where all variates are subject to error

1. *Introduction.* — When it is desired to fit a straight line, $y = a_1 x + b_1$, to $m > 2$ observed points, (x_i, y_i), the method usually employed is that of choosing a_1 and b_1 so that they minimize the sum of the squares of the residuals of the y's, *i.e.* $\Sigma (a_1 x_i + b_1 - y_i)^2$ is a minimum. It is well known that, unless all observed points lie on a line, there are two such lines of « best, fit » (1), one obtained by choosing x as the independent variable, the other obtained by choosing y as the independent variable.

A more difficult problem arises if both variates are subject to error. Here there is no reason to prefer one of the above described lines to the other. As an example, the problem of deriving the law of demand for sugar requires that we find the equation of the « best fitting » straight line (2), an instance in which we are faced with the dilemma of deciding whether the price of sugar or the quantity consumed is the independent variable ; there are two lines of « best fit », one for each of the two possible choices we can make. We know, however, that both the statistics of sugar consumption and of sugar prices are subject to error, and there is, therefore, no reason why we should choose one line to the exclu-

* The present paper develops an idea presented by the author and Alexander Oppenheim to the American Mathematical Society at its meetings in Nashville, Tenn., December 1927. For an abstract of the paper presented there see the « Bulletin of the American Mathematical Society », Vol. 34, 1928, pp. 140-141.

(1) See any elementary text in statistics, for instance, DAVIS and NELSON, « *Elements of Statistics* », 1935, pp. 271-275.

(2) HENRY SCHULTZ, *The Statistical Law of Demand*, « Journal of Political Economy », Vol. 33, No. 6, Dec. 1925, p. 578.

sion of the other. Moreover, as we shall see neither line is the « best » to represent the demand for sugar. A more general problem arises when it is desired to determine the demand for gasoline as a function of price, purchasing power, road mileage, and taxes (1). Here the last two are practically free of error, but the three-demand, price, and purchasing power - all involve errors of comparable magnitude. In fact, in this instance the errors in these three quantities would be correlated due to the omission of factors affecting all three.

From the two examples cited here it seems to be quite clear, therefore, that we should seek some other criterion of fit. We could with R. J. Adcock (2) minimize the sum of the squares of the normals from the m points to the required line of « best fit ». Thus in a more recent paper on the same subject L. J. Reed (3) has written : « in practically all cases of observed data x is as subject to variation as y and it, therefore, appears *a priori* that a better fitting straight line would be obtained if we define the word residual as the normal deviation of an observed point from the line (2). This definition assumes that an observed point fails to fall on the line due to an error in both x and y ». Starting from the normal form of the equation of a line, $x \cos \theta + y \sin \theta - p = 0$, Reed then obtained simple equations to determine the θ and p which minimize the sum of the squares of the normal deviates ; that is, which minimize $\Sigma r_i^2 = \Sigma (x_i \cos \theta + y_i \sin \theta - p)^2$. But there is a difficulty which Reed apparently did not appreciate.

Let the line obtained by Reed's method be $x \cos \theta + y \sin \theta - p = 0$. If we change the scale of x, i.e. write $x = \lambda X$, where λ is a constant, the line obtained will be $X \cos \theta' + y \sin \theta' - p' = 0$ or returning to our original scale $x \cos \theta' + y \lambda \sin \theta' - \lambda p' = 0$, and it is not in general true that this line is the

(1) See C. F. Roos, *Dynamic Economics*, 1934, Chapter III, for demand of gasoline.

(2) R. J. Adcock, « Analyst », Vol. 4, 1877, p. 183, and Adcock, *A Problem in Least Squares*. « Analyst », Vol. 5, 1878, pp. 53-54. The formula for a given by Adcock on page 54 is incorrect because of an error made in eliminating b. For another early attempt to solve this problem see Karl Pearson, *On lines and planes of closest fit, to systems of points in space*. « Phil. Mag. », 6 Ser., Vol. 2, Nov. 1901. It seems that Pearson was ignorant of the existence of Adcock's paper.

(3) L. J. Reed, *Fitting Straight Lines*, « Metron », Vol. 1, (3), 1921, pp. 54-61.

same as the first one. It seems necessary, therefore, to attach from the start weights to the observations x and y, which will naturally vary with the scale.

In considering the same problem of minimizing normal deviates as applied to the fitting of a line to the link relatives, $Y = y(t)/y(t-1)$ and $X = x(t)/x(t-1)$, where $y(t)$ is the annual consumption of sugar at the time t and $p(t)$ its price, Henry Schultz has remarked :

« Indeed, if the method is to be applied by direct processes, the further assumption has to be made that the errors of X and Y have a constant ratio. When this ratio is equated to unity that is to say *when the observations in X are given the same weight as the observations in Y* ....., this method has a simple geometrical interpretation » (1), a minimization of distances of the normals from the observed points to the line of fit.

In appraising this remark in the light which will be shed by the analysis about to be presented, the reader should bear in mind that the link relatives Y and X are independent of the initial scales and that when variates not possessing this property are used the normal line need not be the one for which the weights are equal. Moreover, in Schultz' problem the standard deviations of these ratios are nearly equal, and as we shall see in Section 6, this condition is implicitly required when normal deviates are minimized ; in his problem it is the ratio of the errors of the link relatives X and Y and not of the raw variables x and y that is equal to unity.

From the preceeding remarks it appears, therefore, that as advocated years ago by Kummell (2), it would be better to minimize the weighted sum, $\Sigma\, p_i (\delta x_i)^2 + q_i (\delta y_i)^2$, where δx_i and δy_i are the respective errors in x_i and y_i and p_i and q_i are the weights of x and y which enter the same observation equation. To obtain a direct solution some assumption with regard to the weights of the observed quantities must be made. Kummell found it most convenient to suppose that the weights of x and y which

---

(1) HENRY SCHULTZ, *The Statistical Law of Demand*, loc. cit., p. 581.

(2) C. H. KUMMELL, *Reduction of observation equations which contain more than one observed quantity*, « Analyst », Vol. 6, No. 4, July 1879, pp. 97-105. See also M. MERRIMAN, *The determination, by the method of least squares, of the relation between two variables connected by the equation* $Y = AX + B$, *both variables being liable to errors of observation*, « Report of U. S. Coast and Geodetic Survey », 1890, p. 687.

enter the same observation equation have a constant ratio, i.e. $q_i = k\,p_i$, $(i = 1, \ldots, m)$. Then obviously, as he pointed out, for the special case $p_1 \equiv p_2 \equiv \ldots \equiv p_m \equiv 1$ and $k \equiv 1$, this problem reduces to the earlier one proposed by Adcock. Kummell, it may be noted further, considered not merely this problem of fitting a straight line, but also that of fitting any curve, and his work, although his mathematical arguments are loose or even in some instances fallacious, contains practically all the results obtained by later writers.

In a more recent paper, presumably written in ignorance of this author's work, H. S. Uhler (1) shows that if (u_i, v_i) is the point on the line of best fit $ax + by + 1 = 0$ corresponding to the observed point (x_i, y_i), and p and q are constant weights, then the ratio of the errors $(v_i - y_i)/(u_i - x_i)$ is equal to $(b/a)\ (p/q)$ which is independent of i; that is to say, the lines joining the observed points to the corresponding points on the line are parallel. This result is obtained by minimizing

$$W = p\,(u_i - x_i)^2 + q\,(v_i - y_i)^2 \tag{1}$$

with respect to u_i and v_i subject to the condition $au_i + bv_i + 1 = 0$. After Uhler has minimized W with respect to u_i and v_i he has a sufficient number of equations to enable him to eliminate u_i and v_i from W. Upon performing the elimination he obtains an expression of the form $W_m = ((pq)/(qa^2 + pb^2))\ (ax_i + by_i + 1)^2$ which he minimizes with respect to a and b. In order that the lines of correction be normal to the curve it is sufficient that $p = q$ constant. But since Uhler has taken the p_i and q_i to be constants so that necessarily $p_i/q_i =$ constant all these results are contained implicitly in Kummell's formulas (2).

In 1921 Corrado Gini (3) attacked the special case of the general problem, for which (1) the errors δy_i and δx_i in y and x are uncorrelated, that is, $\Sigma\,\delta y_i\,\delta x_i = 0$, and (2) the ratio of the variances of the errors is constant or $K = \Sigma\,\delta x_i^2/\Sigma\,\delta y_i^2$. Briefly, Gini's

---

(1) H. S. Uhler. *Method of Least Squares and Curve Fitting.* « Journal Amer. Opt. Society », Vol. 7, 1923, pp. 1043-1066.

(2) Kummell, *l. c.* p. 99, (5), (*a*) and p. 101, (*e*).

(3) Corrado Gini, *Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali*, « Metron », Vol. 1, No. 3, 1921, pp. 63-82.

important method consists in modifying the ordinary least squares' solution by a suitable correction as a function of the error of the independent variable. For the first assumption he shows that if x is taken as the independent variable, the slope of the line of best fit β' is $< \beta$, where β is the value obtained by supposing x exact, and if y is taken as the independent variable, the slope of the line of best fit is $\frac{1}{\beta''} < \frac{1}{\beta}$. Then using the second, he obtains values of β involving k, a covariance, and several variances. In particular, for $k = 1$, Gini's general solution reduces to that of Kummell and the later rediscovery of Pearson.

In two subsequent papers G. Pietra has extended Gini's methodology to parabolic functions and more general types defined by Taylor's series (1). For instance, assuming for large numbers of observations, $\Sigma \delta x_i^2 / \Sigma \delta y_i^2 = K$ a constant, $\Sigma \delta x_i^m = 0$ and $\Sigma \delta y_i^m = 0$ for m odd, and $\Sigma \delta x_i \delta y_i = 0$, he shows that, for the parabola $y = \alpha + \beta x + \lambda x^2$, the parameters are given by the equations,

$$\lambda = \frac{\Sigma x_i^2 y_i - \dfrac{\Sigma x_i y_i}{\Sigma x_i^2 - \Sigma \delta x_i^2} \cdot \Sigma x_i^3}{\Sigma x_i^4 - \dfrac{(\Sigma x_i^2)^2}{n} - 2 \Sigma \delta x_i^2 (2 \Sigma x_i^2 - \Sigma \delta x_i^2) - \dfrac{(\Sigma x_i^3)^2}{\Sigma x_i^2 - \Sigma \delta x_i^2}}$$

$$\beta = \frac{\Sigma x_i y_i - \gamma \Sigma x_i^3}{\Sigma x_i^2 - \Sigma \delta x_i^3} \qquad (2)$$

$$\alpha = -\gamma \frac{\Sigma x_i^2 - \Sigma \delta x_i^2}{n}$$

where x_i and y_i refer to the observed values of x and y as measured from their respective means.

It should be recognized, however, that there are various practical cases involving relatively small numbers of observations for which the assumptions of Gini and Pietra do not hold. For instance if the first n of the observed points $(u_i\, v_i)$ lie below the line of fit and the last n lie above, as might well happen for

---

(1) G. PIETRA, *Interpolating Plane Curves*, « Metron », Vol. 3, No. 3-4, 1924, pp. 311-328. See also, G. PIETRA, *Dell'interpolazione parabolica nel caso in cui entrambi i valori delle variabili sono affetti da errori accidentali*. « Metron », Vol. 9, Nos. 3-4, 1932, pp. 77-86.

small numbers of observation, $\Sigma x_i \delta y_i$ is the sum of positive terms and cannot vanish *unless* the origin is at the centroid, and similarly for $\Sigma y_i \delta x_i$. These remarks apply to the papers by both Gini and Pietra. In the latter's paper there is the additional assumption :

$$(3) \qquad \begin{aligned} \Sigma \delta x_i^m &= 0 \qquad \text{for } m \text{ odd} \\ \Sigma \delta y_i^m &= 0 \qquad \text{and } n \text{ large} \end{aligned}$$

Since we may take any number of equations (3), we may in particular choose $m = n - 1$. We may then order our observations so that δx_n is not zero and divide by it in the first set. We will then be able to solve for the ratios $\delta x_i / \delta x_n$ once and for all for any set of data unless $\delta x_i / \delta x_n = \delta x_s / \delta x_n$, since otherwise the functional determinant, which happens to be of the Vandermonde type, will not vanish. Therefore, the exception must hold and consequently Pietra's assumption is valid only for a symmetrical distribution of the errors about the line of fit.

The problem has been attacked from a different point of view by E. C. Rhodes (1), who takes the equation of the fitting line, or more generally the fitting plane, to be of the form $m_1 x/u_1 + m_2 y/u_2 + \ldots + m_n z/u_n - p = 0$, where $m_1^2 + m_2^2 + \ldots + m_n^2 = 1$ and $u_1 \ldots u_n$ are quantities measured in the same units as $x, y, \ldots z$, respectively, and which reduce the last to a common base. He then introduces $(n - 1)$ parameters $t_1, \ldots t_{n-1}$ determining the $x, y, \ldots z$ by the equations

$$\begin{aligned} x/u_1 &= p\, m_1 + a_{11}\, t_1 + \ldots + a_{1\,n-1}\, t_{n-1} \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ z/u_n &= p\, m_n + a_{m1}\, t_1 + \ldots a_{mm-1}\, t_{m-1} \end{aligned}$$

where the $a's$ are unknown functions of the $m's$ such that

$$\begin{aligned} m_1 a_{11} + m_2 a_{21} + \ldots m_n a_{n1} &= 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ m_1 a_{1\,n-1} + m_2 a_{2\,n-1} + \ldots + m_n a_{n\,n-1} &= 0 . \end{aligned}$$

---

(1) E. C. Rhodes. *On lines and planes of closest fit.* « Phil. Mag. », 7 Ser., Vol. 3, (1927), pp. 357-364. See also, H. Glauert. *Best linear relationship connecting any number of variables.* « Phil. Mag. », (1925), p. 205.

He next supposes that the most likely t's will be those which maximize the chance of the measurement made lying between the limits x , $x + dx$, . . . z , $z + dz$. This chance is

$$F = (1/(2\pi)^{n/2}\, s_1, \ldots, s_n)\ (\text{exponential})\ \{-1/2\ E\ (d\,x/u_1) \ldots (d\,z/u_n)\}$$

where $E = (x/u_1 - p\ m_1 - a_{11}\ t_1 - \ldots - a_{1\,n-1}\ t_{n-1})^2/s_1^2 + \ldots +$

$$+ (z/u_n - p\ m_n - a_{n1}\ t_1 - \ldots - a_{n\,n-1}\ t_{n-1})^2/s_n^2,$$

and where the quantities s_1 , . . . , s_n are defined to be the standard deviations of the errors involved in measuring the x/u_1 , . . . , z/u_n , respectively.

The equations resulting from differentiating F with respect to t_1 , . . . , t_{n-1} respectively, enabled Rhodes to eliminate the t's and discover an expression which involves only the m's and p's as variables. He then differentiated the expression so obtained with respect to p and found that the best fitting « plane » goes thru the centroid x_1 , y_1 , . . . , z_1 . He then transfered his origin to the centroid and introduced the notation $\sigma^2_x = \Sigma\,(x - x_1)^2$, $r_{xy} = \Sigma\,(x - x_1)\,(y - y_1)/N\,\sigma_x\,\sigma_y$ of the correlational calculus and substituted these symbols in F. By minimizing the resulting expression with respect to m_1 , . . . , m_n he deduced the equations for the determination of m_1 , . . . , m_n . These equations, of course, involve the standard deviations s_1 , . . . , s_n and the quantities u_1 , . . . , u_n which reduce the x , y , . . . , z to a common base. In particular, if we choose all of the u's and the s's to be equal, Rhodes' equations reduce to those obtained by Pearson (1) for minimizing normal distances.

It is thus seen that there exist various methods for fitting lines and planes in cases where all variables are subject to error, and that each scheme depends upon the choice of a particular coordinate system. In general, we may obtain satisfactory lines (or planes) of fit by any one of a half dozen or more methods and yet if we should change our system of coordinates, we would obtain different lines. It would seem, therefore, that something is obviously lacking, for the problem should be capable of a unique solution regardless of the coordinate system chosen. It is this phase of the problem that we consider in the present paper.

---

(1) KARL PEARSON. *On Lines and Planes of Closest Fit to Systems of points in Space*, loc. cit.

2. *The form of the expression to be minimized.* — In seeking for an invariant criterion of fit it is quite important that we choose the proper function to be minimized, since as we have already pointed out this function must be independent of the choice of axes. It is certain that this function must have as arguments the observed quantities (x_i, y_i) and the coordinates a, b, and c of the line of fit, and we may assume with enough generality that the expression is of the form $U = \Sigma\, p_i\, f\,(x_i, y_i, a, b, c)$ where p_i are for the present merely numbers — later to be identified as the weights belong to observations (x_i, y_i) — and a, b, and c are the coordinates of the line of fit, $ax + by + c = 0$. As we shall now see the condition that U be invariant for all changes of axes is sufficient to determine sufficiently completely its form.

Let us first apply homogeneous strain, *i.e.* write $x = \lambda X, y = \mu Y$ where λ and μ are arbitrary parameters. Then the observations (x_i, y_i) become X_i, Y_i and the line of fit

$$a\,x + b\,y + c = 0 \tag{4}$$

becomes $AX + BY + C = 0$ where $A = \lambda\, a, B = \mu\, b, C = c$. Since by hypothesis the line of fit in one coordinate system must transform into the line of fit in a second, if U is to be invariant, we must have

$$\Sigma\, p_i\, f\,(\lambda\, X_i, \mu\, Y_i, \lambda^{-1} A, \mu^{-1} B, C) = \Sigma\, p_i\, f\,(X_i, Y_i, A, B, C)$$

for all choices of the numbers p_i; and hence

$$f\,(\lambda\, X, \mu\, Y, \lambda^{-1}\, A, \mu^{-1}\, B, C) = f\,(X, Y, A, B, C)$$

for all λ and all μ. If we now follow a well-known procedure of differentiating partially with respect to λ and then putting $\lambda = \mu = 1$, we obtain

$$X\,(\partial f/\partial X) - A\,(\partial f/\partial A) = 0, \tag{5}$$

where for convenience in notation we have suppressed the subscript i. Similarly by differentiating with respect to μ and setting $\lambda = \mu = 1$ we obtain

$$Y\,(\partial f/\partial Y) - B\,(\partial f/\partial B) = 0. \tag{6}$$

The general solution of these partial differential equations (5) and (6) is

$$f = \varphi\,\{A\,X + B\,Y + F\,(C)\} \tag{7}$$

where $F(C)$ is arbitrary (1). The equation (7) is merely a necessary condition but as may readily be verified, (7) is actually invariant for the transformation in question, namely homogenous strain.

Let us next make use of the condition that f be invariant for translation. For this write $X = x + \alpha$, $Y = y + \beta$, and then if $f = \varphi$ is to be invariant, we must obviously have

$$(8)\ \varphi\{a x + b y + F(c + a\alpha + b\beta)\} = \varphi\{a x + b y + a\alpha + b\beta + F(c)\}$$

for all α and all β. If we differentiate the whole expression (8) partially with respect to α and then put $\alpha = \beta = 0$, we obtain the condition $(\partial\varphi/\partial F)\, F' - \varphi' = 0$ where primes denote derivatives. Now since, for $\alpha = \beta = 0$, F is a function of c alone, this condition can be written $(\partial\varphi/\partial c) - \varphi' = 0$. In order that this last partial differential equation be satisfied it is necessary and sufficient that $F(c) = c$. We can, therefore, say that $f(x_i\, y_i\, a, b, c)$ must be of the form

$$(9)\qquad f = \varphi(a x_i + b y_i + c)$$

where φ is an arbitrary function ; and we verify by direct substitution that this expression is invariant for both homogenous strain and translation.

We know, moreover, that φ must be homogenous in a, b, and c, since the line $\lambda ax + \lambda by + \lambda c = 0$ is the same as the line (4), a condition which requires that (9) assume the form

$$(10)\qquad f = (a x_i + b y_i + c)^n$$

where n is arbitrary. Thus we have found for the set of points (x_i, y_i), the associated numbers p_i and the line (4), the most general function $\Sigma\, p_i f(x_i, y_i, a, b, c)$ which remains invariant under translation and homogenous strain. In addition the expression (10) is actually invariant for rotation when the a, b, and c are parameters defined by the fitting line $ax + by + c = 0$.

3. *Condition that U be a function of distances.* — At this stage our problem of fitting a line requires the finding of a line (4) associated with points $(x_i\, y_i)$ and numbers p_i such that $U = \Sigma\, p_i f = \Sigma\, p_i\, (ax_i + by_i + c)^n$ is a minimum. Now, clearly if U is to

(1) E. Goursat. *Cours d'analyse mathematique*, Vol. 2, p. 559.

have a minimum, there must be some relation between a, b and c, for, as matters stand, there would be three normal equations to determine the two ratios a/c and b/c. In choosing this relation we observe that, at least in practical cases, it seems desirable for $ax_i + by_i + c$ to be proportional to the distance between $(x_i\ y_i)$ and the associated point (u_i, v_i) on the line (4), the factor of proportionality being independent of a, b, c and i. Such a relation must, of course, be invariant for all changes of axes and, therefore, can not be of the form $\theta\ (a, b, c) = \text{constant}$ as may readily be verified; the function θ must also involve some parameter and preferably one associated with the accuracy of the data.

We could at this point make various assumptions, among which would be the hypothesis $K = \Sigma\, \delta\, y_i^2 / \Sigma\, \delta\, x_i^2$ of Gini. Similarly we could choose still different metrics which reconcile other methods of fitting with the results so far obtained. In the following discussion we prefer to assume that the line joining (x_i, y_i) to (u_i, v_i) makes an angle α independent of i with the axis of x, *i.e.* $\delta\, y_i / \delta\, x_i = \tan \alpha = k$, a constant. Then, the condition that $ax_i + by_i + c$ be proportional to the distance $r_i = \{ (u_i - x_i)^2 + (v_i - y_i)^2 \}^{\frac{1}{2}}$, the factor of proportionality being independent of a, b, c and i, requires only that $a \cos \alpha + b \sin \alpha$ be independent of a, b, c and i; for $u_i = x_i + r_i \cos \alpha$, $v_i = y_i + r_i \sin \alpha$ and $au_i + bv_1 + c = 0$, so that $(ax_i + by_i + c) + r_i (a \cos \alpha + b \sin \alpha) = 0$, whence it follows that

(11) $a \cos \alpha + b \sin \alpha$ must be independent of a, b, c and i.

We may for definiteness take

(12) $a \cos \alpha + b \sin \alpha = 1$,

although all that we really require is that

(13) $\cos \alpha\, \delta\, a + \sin \alpha\, \delta\, b = 0$,

where δa and δb are respectively the variations of a and b. It should be noticed, however, that in a different coordinate system (12) will assume some such form as $A \cos \alpha' + B \sin \alpha' = f(\alpha')$, where again $A \cos \alpha' + B \sin \alpha'$ is independent of A, B, C and i, which is the essential condition. For example, if we assume (12) in the coordinate system $\{ x, y \}$ then $ax_i + by_i + c$ actually represents the distance r_i, but in any other coordinate system $\{ X, Y \}$ the corresponding expression $AX_i + BY_i + C$ represents the distance R_i corresponding to r_i multiplied by a function which

is independent of A, B, C and i. Consequently the line of fit $AX + BY + C = 0$ does actually correspond to $ax + by + c = 0$; *i.e.* we can get (4) by changing to the axes x, y, in $AX + BY + C = 0$.

We will assume the condition to be in the form (11) in the work which follows.

4. *The Problem of Fitting a Straight Line.* — If we assume the condition (11) then we will necessarily want U to be the sum of positive terms in order to be assured of a good fit. We, therefore, define U by the formula

$$U = \Sigma\, p_i \mid a\, x_i + b\, y_i + c \mid^n \tag{14}$$

since by hypothesis the weights p_i are all positive, and $f = (ax_i + + by_i + c)^n$ by (5). We are at liberty to take n as we please; thus in particular if we take it as 2, the quantity U defined by (14) becomes some function independent of a, b, c, and i, times the weighted sum of the squares of the distances r_i between the observed point (x_i, y_i) and the corrected point (u_i, v_i) on the line (1). It is usually undesirable to take $n > 2$ since this gives emphesis to large errors, and for distributions that commonly occur this, to say the least, is questionable procedure. If we minimize U with respect to a, b and c which satisfy (12) or the equivalent (13) we obtain values of a, b and c which minimize the weighted sum of the squares of the r_i.

Our problem now is to reduce

$$U = \Sigma\, p_i\, (a\, x_i + b\, y_i + c)^2 \tag{15}$$

to a minimum when a, b, and c are subject to the condition (13).

The conditions for a minimum are

$$\Sigma\, p_i\, [(a\, x_i + b\, y_i + c)\, x_i\, \delta\, a + (a\, x_i + b\, y_i + c)\, y_i\, \delta\, b + + (a\, x_i + b\, y_i + c)\, \delta\, c] = 0\,,$$

$$\cos\, \alpha\, \delta\, a + \sin\, \alpha\, \delta\, b = 0\,.$$

These equations of variation lead to the conditions

$$\Sigma\, p_i\, x_i\, (a\, x_i + b\, y_i + c)\, \sin\, \alpha = \Sigma\, p_i\, y_i\, (a\, x_i + b\, y_i + c) \cos\, \alpha \tag{16}$$

$$\Sigma\, p_i\, (a\, x_i + b\, y_i + c) = 0 \tag{17}$$

In as much as the ratios a/c and b/c determine uniquely the line (1), equations (16) and (17) are sufficient in number. In

particular, (17) shows us that the line of fit passes through the weighted mean center.

Let us by analogy call the equations (16) and (17) the general normal equations since when $\alpha = \pi/2$, or $\alpha = 0$ they reduce to the equations usually called normal equations.

Simplifying notations we use

$$[p\,x^2] = \Sigma\, p_i\, x_i^2\,,\ [p\,x\,y] = \Sigma\, p_i\, x_i\, y_i\,,\ [p\,y^2] = \Sigma\, p_i\, y_i^2$$
$$[p\,x] = \Sigma\, p_i\, x_i\,,\ [p\,y] = \Sigma\, p_i\, y_i\,,\quad \text{and}\quad [p] = \Sigma\, p_i\,,$$

and then write equations (16) and (17) as

$$(18)\quad \left.\begin{aligned}(\sin\alpha\,[p\,x^2] - \cos\alpha\,[p\,x\,y])\,a + (\sin\alpha\,[p\,x\,y] - \cos\alpha\,[p\,y^2])\,b + \\ + (\sin\alpha\,[p\,x] - \cos\alpha\,[p\,y])\,c = 0\end{aligned}\right.$$

$$(19)\quad [p\,x]\,a + [p\,y]\,b + [p]\,c = 0\,.$$

By using $\tan\alpha = k$ we may put (18) in the simplified form

$$(20)\quad (k\,[p\,x^2] - [p\,x\,y])\,a + (k\,[p\,x\,y] - [p\,y^2])\,b + (k\,[p\,x] - [p\,y])\,c = 0\,.$$

5. *Solution of normal equations:* Equations (18) and (19) are sufficiently simple to be used readily in computation as they stand, but it will be interesting and instructive actually to find a solution for the ratios a/c and b/c for the general case. Solving (19) for a/c and substituting the value so obtained in (18) we may write

$$(21)\quad \frac{b}{c} = \frac{([p\,x\,y] - k\,[p\,x^2])\,[p] - ([p\,y] - k\,[p\,x])\,[p\,x]}{(k\,[p\,x^2] - [p\,x\,y])\,[p\,y] - (k\,[p\,x\,y] - [p\,y^2])\,[p\,x]}$$

and again if we solve (19) for b/c and substitute this value in (18), we obtain

$$(22)\quad \frac{a}{c} = \frac{(k\,[p\,x\,y] - [p\,y^2])\,[p] - (k\,[p\,x] - [p\,y])\,[p\,y]}{(k\,[p\,x^2] - [p\,x\,y])\,[p\,y] - (k\,[p\,x\,y] - [p\,y^2])\,[p\,x]}$$

Let $(X\,,\ Y)$ be the running coordinates of the line of best fit. Then the equation of the line of best fit is

$$(23)\quad \begin{aligned}&\{(k\,[p\,x\,y] - [p\,y^2])\,[p] - (k\,[p\,x] - [p\,y])\,[p\,y]\}\,X + \\ &+ \{([p\,x\,y] - k\,[p\,x^2])\,[p] - ([p\,y] - k\,[p\,x])\,[p\,x]\}\,Y + \\ &+ (k\,[p\,x^2] - [p\,x\,y])\,[p\,y] - (k\,[p\,x\,y] - [p\,y^2])\,[p\,x] = 0\end{aligned}$$

We have seen that for our problem we may choose any point as the origin of coordinates ; but it happens that the choice of the weighted mean centre materially simplifies calculations. When such a choice is made, $[px]$ and $[py]$ are both zero, and equation (19) becomes

$$(24) \qquad -([p\,y^2] - k\,[p\,x\,y])\,X + ([p\,x\,y] - k\,[p\,x^2])\,Y = 0$$

This is a form which is sufficiently simple to appeal to any calculator.

6. *Important special cases.* — There are several important special cases of our general problem which are worthy of individual mention. For example if the x's are perfect, $\tan \alpha = \infty$, or the equivalent $\sin \alpha = 1$, $\cos \alpha = 0$. From (8) we see then that the conditions are

$$[p\,x^2]\,a + [p\,x\,y]\,b + [p\,x]\,c = 0$$
$$[p\,x\,]\,a + [p\,y]\;\; b + [p]\;\; c = 0$$

which are the equations usually obtained by minimizing the sum of the squares of the residuals of the y's. Again if the y's are supposed perfect, $\tan \alpha = 0$, or the equivalent $\sin \alpha = 0$, $\cos \alpha = 1$. For this case equations (18) and (19) reduce to

$$[p\,x\,y]\,a + [p\,y^2]\,b + [p\,y]\,c = 0$$
$$[p\,x]\;\; a + [p\,y]\;\, b + [p]\;\; c = 0$$

and again these are the usual equations for this criterion.

If we take $\tan \alpha = b/a$, we see without difficulty that equation (12) becomes $(a^2 + b^2)^{\frac{1}{2}} = 1$. This is equivalent to saying that $a = \cos \theta$, $b = \sin \theta$ and $c = -p$, where θ is the angle the line $a\,X + b\,Y + c = 0$ makes with the X axis and p is the distance from the origin to the line. We are thus led to the special case discussed by Reed for which we minimize $\Sigma\,[(x_i \cos \theta + y_i \sin \theta - p)^2]$. In view of these considerations, it seems to us that to say normal distances should be minimized is, in general, somewhat equivalent to assuming in advance a solution of the problem. As an example suppose it is known that the precisions in x and y have a ratio of unity. Then α must be $-\pi/4$; and for this case the minimizing of normal distances would be correct only if b/a was unity. Therefore, normal distances should be minimized only

in the very special case for which the x's and y's are both independent of scale factors and the precisions of, say, x/σ_x and y/σ_y are equal : it would in general, be artificial to say that whenever x and y are both subject to error, the ratio of their precisions is equal to the negative reciprocal of the slope of the line of best fit.

In using Pearson's method of fit, *i . e .*, the minimizing of normal deviates, the work is simplified if the hypothesis is made that each variate is reduced to units of its standard deviation, so that each has unit variance. But graphical curve-fitters have not always appreciated this special assumption ; for we find workers in the U. S. Department of Agriculture and elsewhere merely expressing each variate as a ratio to its mean — which of course is something quite different — and fitting free-hand curves. In such latter instances any line of regression may be determined, the particular one obtained (1) depending upon what the standard deviations of the series are. It is the effect of the scale factor leading to different lines of fit that causes uncertainty and disagreement among graphical curve-fitters.

Briefly if lines are to be fitted free-hand, each variate should be expressed in units of its standard deviation and then plotted on equal scales, or the worker should give definite reasons for chosing the scales or systems of weights.

It is true that for the cases where the criterion of minimizing normal deviates has been used, the resulting line of « best fit » has seemed satisfactory and has been more gratifying than either the line for which the x's are assumed perfect or the one for which the y's are assumed perfect, but should we not expect this ? The point that we make is that, although the line obtained by minimizing the sum of the squares of the normal distances gives a better line than either of the other two lines referred to in the preceding sentence, this line is not always the correct one to be used ; *i . e .* the hypothesis itself is at serious fault.

7. Extension to the Fitting of Planes and Hyper-planes. — As we have already intimated, our analysis generalizes immediately to any number of dimensions. More precisely if we desire to fit a plane, we minimize the invariant expression $\Sigma p_i (ax_i + by_i + cz_i + d)^2$

(2) See HERBERT E. JONES. *Some Geometrical Considerations in the General Theory of Fitting Lines and Planes.* « Metron », Vol. XIII, n. 1 February, 1937, p. 21-30.

subject to the condition $am_1 + bm_2 + cm_3 = \varphi$ independent of a, b, c, d and i, where m_1, m_2 and m_3 are the direction cosines of the line joining the observed point (x_i, y_i, z_i) to the corrected point (u_i, v_i, w_i) on the plane of fit $ax + by + cz + d = 0$. Then, by the ordinary rules of the calculus of maxima and minima we have

$$(25)\left\{\begin{array}{c}\Sigma\, p_i\,(a\,x_i + b\,y_i + c\,z_i + d)\,x_i/m_1 = \Sigma\, p_i\,(a\,x_i + b\,y_i + \\ + c\,z_i + d)\,y_i/m_2 = \Sigma\, p_i\,(a\,x_i + b\,y_i + c\,z_i + d)\,z_i/m_3\end{array}\right.$$

$$(26)\qquad \Sigma\, p_i\,(a\,x_i + b\,y_i + c\,z_i + d) = 0$$

Equation (26) shows that the plane passes through the weighted mean point, which to simplify matters will be taken as the origin. If we adopt our previous notation, we may write:

$$(27)\left\{\begin{array}{c}\{m_1\,[p\,x\,y] - m_2\,[p\,x^2]\}\,a + \{m_1\,[p\,y^2] - m_2\,[p\,y\,x]\}\,b + \\ + \{m_1\,[p\,z\,y] - m_2\,[p\,z\,x]\}\,c = 0\end{array}\right.$$

$$(28)\left\{\begin{array}{c}\{m_2\,[p\,x\,z] - m_3\,[p\,x\,y]\}\,a + \{m_2\,[p\,y\,z] - m_3\,[p\,y^2]\}\,b + \\ m_2\,[p\,z^2] - m_3\,[p\,z\,y]\,c = 0\end{array}\right.$$

where the origin of coordinates is now at the weighted mean. These equations enable us to solve for the ratios a/c and b/c which completely determine the plane of best fit. We may, in fact, write the equation of this plane when the errors are defined by $\delta x_i/m_i = \delta y_i/m_3 = \delta z_i/m_i$ as the following:

$$(29)\qquad \begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix} U - \begin{vmatrix} A_1 & C_1 \\ A_2 & C_2 \end{vmatrix} V + \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix} W = 0$$

where U, V and W are running coordinates referred to the mean as origin; A_1 and A_2 are the coefficients of a in (27) and (28), respectively; B_1 and B_2 are the coefficients of b and C_1 and C_2 are the coefficients of c.

It is very simple to obtain equations corresponding to (27) and (28) for hyper-planes; in fact to obtain additional terms corresponding to those of (27) and (28), we need only permute the m's in the order $1, \ldots, n$ and the second letters of the symbols [ ] in the order $X, Y, Z, \ldots$

8. *The Choice of Other Metrics.* — It has been said that other metrics can be chosen and different lines of « best » fit obtained. Obviously some discussion of these variants of the present theory is indicated here but since there is no end to such choices, we shall merely choose one important other metric to show how the theory must be modified. For this consider the metric employed by Gini. Here we have $\Sigma \delta y_i^2 / \Sigma \delta x_i^2 = K$, or for the line $ax + by + c = 0$, the condition

$$K = \Sigma\left(y_i + \frac{a}{b} x_i + \frac{c}{b}\right)^2 / \Sigma\left(x_i + \frac{b}{a} y_i + \frac{c}{a}\right)^2$$

from which we readily deduce

$$K = a^2/b^2$$

corresponding to our relationship $\cos \alpha \, \delta a + \sin \alpha \, \delta b = 0$.

The conditions for a maximum are then :

$$\Sigma p_i \{(a x_i + b y_i + c) x_i \delta a + (a x_i + b x_i + c) y_i \delta b + (a x_i + b y_i + c) \delta c\} = 0$$

and

$$a \delta a = K b \delta b$$

From the last, we have $\delta a = K \frac{b}{a} \delta b$ or since $a^2/b^2 = K$, we obtain

$$\delta a = \pm \sqrt{K} \, \delta b .$$

A substitution in the first equation of condition then leads quickly to the general normal equations. There are, it will be observed, two lines of fit, one being the best and the other the worst.

9. *Some Applications.* — The applications of the present theory are, of course, legion, and all we can hope to do here is to indicate briefly how they work out in practice.

Obviously the first consideration is the *a priori* determination of the relative errors in the variates. This, of course, usually involves some arbitrary assumption. In the linear two-dimensional case mathematical curve-fitters generally assume the ratios to be zero or infinity and proceed to determine one or the other of the regression lines ; exponents of the graphical method, on the other hand, have been observed by the author to minimize

normal deviates (1). In view of these considerations we should not hesitate in the general case to determine our ratios from *a priori* considerations. There are, moreover, some cases in which the relative importance of the errors is known fairly well. Suppose, for instance, that we wish to determine a regression between wholesale prices of cotton goods and farm prices of cotton. We know that over such short intervals of time as given by Table 1, y, the price of cotton goods, is equal to a weighted sum of X, the price of cotton at the farm, Z the average hourly earnings of labor used in cotton manufacturing, and E which is composed of fixed charges and profits or erratic elements; thus

$$Y = a_0 X + a_1 Z + a_2 E$$

For the data of Table I the simple correlation between Y and Z is + .5195 and that between X and Z is + .7764. Both correlations are positive so that we may take

$$-K = \frac{\delta\,(y/\sigma_y)}{\delta\,(x/\sigma_x)} = + \frac{(.\,5195)^2}{(.\,7764)^2} = + \,.\,44$$

TABLE I.

Cost and price in the cotton goods industry.

DATE	Wholesale Prices of Cotton Goods Semi-Annual Av.	Farm Prices of Cotton Goods Semi-Annual Av.	Av. Hourly Earnings of Labor Semi-Annual Av.
	y	x	z
1928 January	153.7	100.1	42.5
1928 July	151.0	100.5	43.2
1929 January	149.7	99.3	42.8
1929 July	140.0	97.7	42.2
1930 January	119.3	90.7	41.9
1930 July	84.2	78.8	41.2
1931 January	74.2	71.0	39.1
1931 July	51.3	61.0	38.0
1932 January	44.5	54.7	36.5
1932 July	48.3	53.6	33.3
1933 January	53.3	54.2	32.9
1933 July	64.7	87.4	43.1
1934 January	81.2	87.3	45.9
1934 July	96.5	86.0	47.0
1935 January	94.8	83.1	46.8
1935 July	86.7	84.1	46.1

(1) The reason for this tendency is, of course, psychological; the eye seeks the shortest distance to the curve of fit and this is the normal.

We, therefore, transform X and Y so that

$$k = \frac{X - (\text{mean of } X)}{\sigma_x}$$

$$k = \frac{Y - (\text{mean of } Y)}{\sigma_y}$$

and substitute directly in the last formula of section 5. If all observations are given the same weight, $p_i = 1$, the line of best fit is

$$-(\Sigma\ y_i^2 + .44\ \Sigma\ x_i\ y_i)\ x + (\Sigma\ x_i\ y_i + .44\ \Sigma\ x_i^2)\ y = 0$$

or more specifically

$$-1.3953\ x + 1.3384\ y = 0$$

If one of the correlation coefficients r_{yz} or r_{xz} had been negative, the number k would have been positive. Similar considerations enable us to determine completely the signs of the direction cosines in the more general case of fitting a hyperplane.

Another particularly interesting and common case arises when we attempt to fit polynomials, for instance,

$$y = a_0 + a_1\ x + a_2\ x^2$$

Suppose that the average $\delta\ x_i = -.3$, and the average $\delta\ y_i = +.2$. The average $\delta\ (x_i^2)$ will be $(-.3)^2 = +.09$, so that the direction cosines are given by

$$\delta\ x_i/m_1 = \delta\ y_i/m_2 = \delta\ (x_i^2)/m_3$$

$$m_1^2 + m_2^2 + m_3^2 = 1$$

and are determined uniquely, in absolute value and in sign. In fact, the method readily generalize to Taylor series or other expansions in which the coefficients appear linearly.

10. *The Errors.* — There is no difficulty in obtaining the standard errors of the coefficients or of the functions themselves, the classical methods of Pearson and others yielding them readily for the particular problem, $k = \tan \alpha$, under consideration. However, the distributions of the regression coefficients in the sense of R. A. Fisher are not so easy to obtain. They remain at this stage a challenge to the mathematical statistician.

HERBERT E. JONES

Some Geometrical Considerations in the General Theory of Fitting Lines and Planes

In a recent article Roos (1) has described the various methods for fitting lines and planes when all variables are subject to error and pointed out that these methods are not independent of the choice of the coordinate system. He has then unified the general theory of series expansions in a single expression, invariant to motion and homogeneous strain, which includes all the previous solutions of the problem as special cases. It is the purpose of the present paper to give a geometrical interpretation of Roos' general solution and some of the special cases.

For this we note that Karl Pearson (2) has shown the geometrical relation between what he has called the plane of "best fit" and the correlation ellipsoid. This plane is obtained by minimizing the sum of the squares of the normal deviates, but is not invariant to homogeneous strain and has very little meaning unless the data are put in terms of their standard deviations. In the latter case Pearson's solution becomes that defined by Roos (3) when the errors $\frac{\delta x}{\sigma_x}$, $\frac{\delta y}{\sigma_y}$ and $\frac{\delta z}{\sigma_z}$ are equal. But Pearson does not place this limitation in his equations (4) and his

(1) C. F. Roos, *A General Invariant Criterion of Fit for Lines and Planes, where all Variates are subject to Error.* "Metron" Vol. XIII, No. 1, pp. 3-20.

(2) K. Pearson, *On Lines and Planes of Closest Fit to Systems of Points in Space,* "Philosophical Magazine" Vol. 2, 6th Series, 1901, pp. 559-572.

(3) *Op. cit.*, pp. 15-16.

(4) Although a cursory reading of Pearson's paper might seem to contradict this statement, his illustrative problems show that he does not define his data in terms of their standard deviations - see pp. 561 et seq., especially pp. 569-572.

plane of " best fit " approaches the ordinary regressions Y on X or X on Y as the ratio σ_y/σ_x approaches zero or infinity; therefore, the title of " best fit " is rather ambiguous, especially in regard to the graphical applications currently being made. However, by extending Pearson's geometrical analysis to Roos' (1) general solution we can readily visualize the various special cases and more fully comprehend the difficulties which arise when we attempt to determine the " best fit ".

For simplicity we shall take the problem of three variables though the formulae may be easily generalized for any number. In this case Roos (5) has defined the plane of fit as,

$$(1) \qquad \begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix} u - \begin{vmatrix} A_1 & C_1 \\ A_2 & C_2 \end{vmatrix} v + \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix} w = 0 ;$$

Where

$$(2) \qquad \begin{aligned} A_1 &= \lambda \sigma_x \sigma_y r_{xy} - \mu \sigma_x^2 , \\ B_1 &= \lambda \sigma_y^2 - \mu \sigma_x \sigma_y r_{xy} , \\ C_1 &= \lambda \sigma_x \sigma_y r_{xy} - \mu \sigma_x \sigma_x r_{xx} , \\ A_2 &= \mu \sigma_x \sigma_z r_{xz} - \nu \sigma_x \sigma_y r_{xy} , \\ B_2 &= \mu \sigma_y \sigma_z r_{yz} - \nu \sigma_y^2 , \\ C_2 &= \mu \sigma_x^2 - \nu \sigma_z \sigma_y r_{zy} ; \end{aligned}$$

the σ's and r's have their usual meaning and λ, μ, ν, are the direction cosines of the lines joining the observed points (x_i, y_i, z_i) and the associated points (u_i, v_i, w_i) on the plane; the weight function, p, is assumed equal to unity.

If we substitute the values given in Eq. (2) in Eq. (1), compute the determinants and simplify, we have

$$(3) \quad (a\lambda + h\mu + g\nu)u + (h\lambda + b\mu + f\nu)v + (g\lambda + f\mu + c\nu)w = 0$$

---

(1) *Op. cit.* p. 17. In equation (2) above the summations of Roos' equation (29) have been replaced by the wellknown statistical constants.

where

$$a = \frac{1 - r^2_{yz}}{\sigma_x^2}, \qquad f = \frac{(r_{xy}\, r_{xz} - r_{yz})}{\sigma_y\, \sigma_z}$$

$$b = \frac{1 - r^2_{xz}}{\sigma_y^2}, \qquad g = \frac{(r_{yx}\, r_{yz} - r_{xz})}{\sigma_x\, \sigma_z} \tag{4}$$

$$c = \frac{1 - r^2_{xy}}{\sigma_z^2}, \qquad h = \frac{(r_{zx}\, r_{zy} - r_{xy})}{\sigma_x\, \sigma_y}$$

Equation (3) is immediately recognized as the equation in current coordinates (u, v, w) of the diametral plane of the quadric surface

$$a\,x^2 + b\,y^2 + c\,z^2 + 2\,f\,y\,z + 2\,g\,x\,z + 2\,h\,x\,y = \text{constant} \tag{5}$$

that bisects the system of parallel chords whose direction cosines are λ, μ, ν. If we substitute for the values, a, b, c, etc., the more common notation involving R_{pq}, the co-factor of the p^{th} row and the q^{th} column in the determinant

$$R = \begin{vmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{vmatrix}$$

equation (5) becomes

$$R_{11}\frac{x^2}{\sigma_x^2} + R_{22}\frac{y^2}{\sigma_y^2} + R_{33}\frac{z^2}{\sigma_z^2} + \tag{6}$$

$$+ 2\,R_{23}\frac{y\,z}{\sigma_y\,\sigma_z} + 2\,R_{13}\frac{x\,z}{\sigma_x\,\sigma_z} + 2\,R_{12}\frac{x\,y}{\sigma_x\,\sigma_y} = \text{constant},$$

which is at once recognized as the correlation ellipsoid.

We may say, therefore, that the plane of fit defined by equation (1) is a diametral plane of the correlation ellipsoid that bisects chords whose direction cosines are the same as those of the lines joining the observed points with associated points on the plane. To extend this conception, we may say, in general, that the plane of fit is a polar plane of the point at infinity, in the direction λ, μ, ν, with respect to the generalized correlation ellipsoid.

The plane of fit most commonly used, is the ordinary least squares solution based on the assumption that the values of the independent variable are exact, all the errors appearing in the dependent variable. We can develop this notion quite easily

with the above concept. If we assume Z to be the dependent variable then we can quickly determine the diametral plane which bisects chords parallel to the Z axis by solving Eq. (6) as a quadratic in Z in terms of the other variables and letting the discriminant equal zero. We then have

$$Z = -\frac{R_{13}}{R_{33}}\frac{\sigma_z}{\sigma_x} X - \frac{R_{23}}{R_{33}}\frac{\sigma_z}{\sigma_y} Y , \tag{7}$$

which is the equation of the plane found by the ordinary least squares solution when Z is the dependent variable. Therefore, we may say in general that the planes determined by the ordinary method of least squares are diametral planes of the generalized correlation ellipsoid bisecting chords parallel to the axis of the dependent variable.

The direction cosines λ , μ , ν , of Eq. (2) for the above simple case are, 1, 0, 0, when X is the dependent variable ; 0, 1, 0, when Y is the dependent variable ; and 0, 0, 1, when Z is the dependent variable and are the limiting values of λ , μ , ν . These three planes form an acute trihedral angle whose vertex is at the center of the correlation ellipsoid and any plane defined by Eq. (1) will lie within this trihedral angle and pass through its vertex, approaching one of the boundary planes as a limit as the errors in any two of the variables become negligible.

The most common special case, developed by Adcock and generalized by Pearson (1), is the case where the sum of the squares of the normal residuals is minimized. As we have seen this plane must be a diametral plane of the correlation ellipsoid and also be perpendicular to the chords it bisects. But this is the definition of a principal plane of the correlation ellipsoid (2). On the other hand, an ellipsoid has three principal planes and the above analysis does not show which one is the plane of " best fit ". Since we are minimizing distances, we would be led to assume that the principal plane perpendicular to the least axis of the ellipsoid would be the plane in question. This could be shown by an analysis of Es. (1) and (6) and was proved by Pearson (3)

---

(1) *Op. cit.*, pp. 559-572.

(2) See any good text in solid analytic geometry, such as, SNYDER and SISAM, *Analytic Geometry of Space*, p. 79.

(3) *Op. cit.*, pp. 567-569.

in a slightly different manner. In passing it might be well to bring out one other point concerning this "normal" plane. In Pearson's analysis a serious difficulty arises when the correlation ellipsoid degenerates into a prolate spheroid, since in that case any plane through the major axis of the ellipsoid satisfies the conditions of minimization. But the diametral planes are uniquely defined by λ, μ, ν, which are independent of the correlation ellipsoid, and so equation (1) is not subject to this limitation.

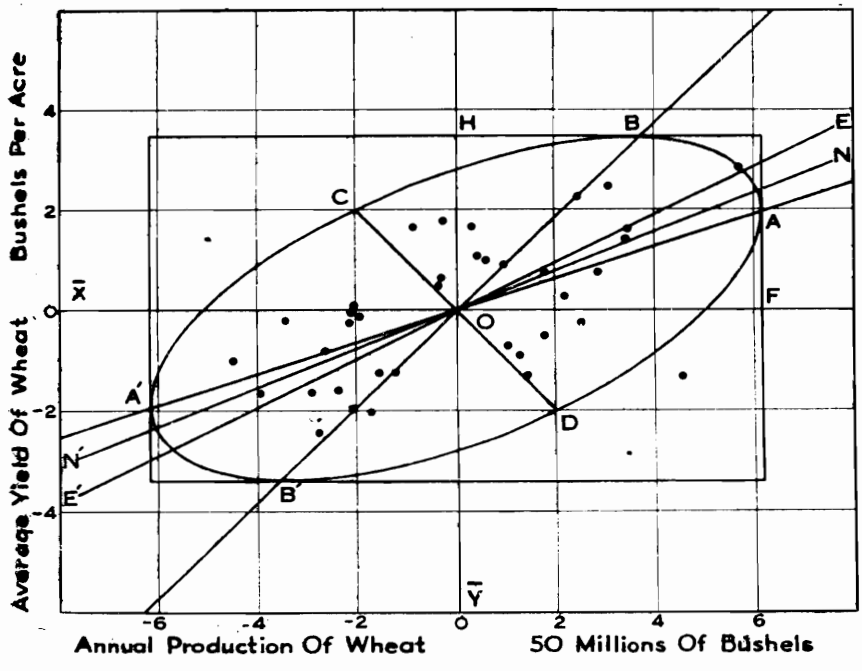


FIG. 1. — Scatter Diagram and Correlation Ellipse for Annual Production of Wheat and Average Yield Per Acre in the United States, 1814-1930.

We can best illustrate the above conceptions by a simple two dimensional case which is shown in Figure 1. This shows a correlation analysis between the annual production of wheat in the United States for the years 1894-1930, and the average yield per acre. The yield is in units of bushels per acre and to make the units comparable for graphical illustration, the production was taken in units of 50 millions of bushels. Deviations from the mean were used in the computations, the mean annual

production being 742.70 millions of bushels, the mean yield being 14.13 bushels per acre. Other units could have been chosen and the effect of the scale factor will be discussed later.

Letting y = yield and x = production, we have

$$\sigma_y = 1.3950\,, \qquad \sigma_x = 2.5286\,,$$

$$r_{xy} = +\,0.5870.$$

The ellipse shown in Figure 1 is the correlation ellipse which should theoretically enclose 95 % of the points, its equation being,

$$(8) \qquad \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\, r_{xy}\, x\, y}{\sigma_x\, \sigma_y} = \chi^2\,(1 - r_{xy}^2)$$

where χ^2 is the frequency function which in this case = 5.9915. In the case of two dimensions equation (1) becomes,

$$(9) \qquad (\sigma_x\, \sigma_y\, r_{xy} - k\, \sigma_x^2)\, y = (\sigma_y^2 - k\, \sigma_x\, \sigma_y\, r_{xy})\, x$$

where $k = \tan \alpha$, α being the angle with the x axis made by the lines joining the observed points with the associated points on the line of fit (1). When $\alpha = \pi/2$, $k = \infty$, we have the common case where y is the only variable subject to error and equation (9) becomes the familiar regression equation of Y on X,

$$(10) \qquad y = r_{xy}\, \frac{\sigma_y}{\sigma_x}\, x\,.$$

Substituting our values from above we have,

$$(11) \qquad y = 0.3238\, x$$

which is line AA' in Figure 1. When x is the only variable subject to error we have $\alpha = 0$, $k = 0$ and equation (9) becomes the familiar regression of X on Y

$$(12) \qquad y = \frac{1}{r_{xy}}\, \frac{\sigma_y}{\sigma_x}\, x$$ and substituting the above values,

$$(13) \qquad y = 1.0642\, x$$

which is shown as line BB' in Figure 1. It follows from the conditions in equations (10) and (12) and from the concept of diameters, that the tangents at points A, A' and B, B' are parallel

(1) C. F. Roos, *op. cit.*, pp. 12-14.

to the Y axis and X axis respectively, forming a circumscribed rectangle about the correlation ellipse. An analysis of equations (8), (10) and (12) shows that the dimensions of this rectangle are $2\,\sigma_x$ in the x direction and $2\,\sigma_y$ in the y direction, the coordinates of the points A and A' being $(\sigma_x\,\chi\,,\ r_{xy}\,\sigma_y\,\chi)$ and $(-\sigma_x\,\chi\,,\ -r_{xy}\,\sigma_y\,\chi)$ respectively and the coordinates of B and B' being $(r_{xy}\,\sigma_x\,\chi\,,\ \sigma_y\,\chi)$ and $(-r_{xy}\,\sigma_x\,\chi\,,\ -\sigma_y\,\chi)$ respectively (1).

Therefore, we see that as r_{xy} varies the shape of the ellipse changes but always remains inscribed in the rectangle whose dimensions are independent of r_{xy}. When r_{xy} equals unity the ellipse degenerates into a straight line coincident with the diagonal of the rectangle; as r_{xy} approaches zero, the ellipse broadens out and its axes rotate about the center, point A approaching F, and B approaching H, and when $r_{xy} = 0$, the axes of the ellipse coincide with the coordinate axes and the regression lines AA' and BB'.

This concept is easily generalized and in the three dimensional case, the correlation ellipsoid is inscribed in a rectangular parallelopiped whose dimensions are $2\,\sigma_x\,\chi\,,\ 2\,\sigma_y\,\chi\,,\ 2\,\sigma_z\,\chi$, the coordinates of the six points of tangency being, $(\pm\,\sigma_x\,\chi\,,\ \pm\ \pm\,\sigma_y\,\chi\,r_{xy}\,,\ \pm\,\sigma_z\,\chi\,r_{xz})$, $(\pm\,\sigma_x\,\chi\,r_{xy}\,,\ \pm\,\sigma_y\,\chi\,,\ \pm\,\sigma_z\,\chi\,r_{yz})$, $(\pm\,\sigma_x\,\chi\,r_{xz}\,,\ \pm\ \pm\,\sigma_y\chi\,r_{yz}\,,\ \pm\,\sigma_z\,\chi)$.

The special case obtained by minimizing the sum of the squares of the normal deviates was shown by Pearson (2) to be the major axis of the correlation ellipse. With the centroid of the system taken as the origin of coordinates this line is defined by

$$(14) \qquad y = \frac{-(\sigma_x^2 - \sigma_y^2) + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\,r_{xy}^2\,\sigma_x^2\,\sigma_y^2}}{2\,r_{xy}\,\sigma_x\,\sigma_y}\,x\,.$$

Substituting the above values in this equation we have,

$$(15) \qquad y = 0.3934\,x$$ which is shown in the figure as line NN'. An analysis of equation (14) shows that the slope of this line will depend on the relative values of σ_x and σ_y.

---

(1) It is interesting to note that in his celebrated paper, *Analyse mathématique sur les probabilités des erreurs de situation d'un point*, " Mémoires de l'Institut de France " IX, 1846, pp. 255-332, Bravais found the equation of the line QB in his analysis of the normal frequency surface for two variables but had no idea it would ever become renowned as a " regression line ".

(2) *Op. cit.*, pp. 565-566.

As the ratio $\sigma_x/\sigma_y \to \infty$ the line NN' will approach AA' in the limit and as $\sigma_x/\sigma_y \to 0$, NN' approaches BB'. If $\sigma_x = \sigma_y$, or if the data are in terms of their standard deviations, the rectangle will become a square and NN' will be the diagonal. But when $\sigma_x > \sigma_y$, NN' will lie below the diagonal while if $\sigma_x < \sigma_y$ it will be above. To illustrate this let us choose 1 million bushels as the unit of annual production instead of 50 millions. We would then have $\sigma_{x'} = 126.4303$ and equation (14) would become

$$y = 0.00648\, x'. \tag{16}$$

This is not equivalent to replacing x with $\frac{x'}{50}$ in equation (15) for in that case we could have.

$$y = 0.3935\, \frac{x'}{50} = 0.00787\, x'\,. \tag{17}$$

On the other hand, if we chose the unit of production to be 500 millions of bushels we have $\sigma_x'' = 0.25286$ and substituting in equation (14),

$$y = 9.2004\ x'' \tag{18}$$

Again this is not equivalent to substituting 10 x'' in equation (15) for then we would have

$$y = 0.3935 \times 10\, x'' = 3.935\, x'' \tag{19}$$

Clearly, then, the title of " best fit " is misleading since this " normal " line may have any slope between AA' and BB' depending upon the scale factor.

For Roos' generalized case in two dimensions, defined by equation (9), we see that the slope of the line of fit depends upon k, the values of k being ∞ when all the errors lie in y, and 0 when all the errors lie in x. Therefore, when it is assumed that both variables are subject to error AA' and BB' become the limiting cases, all other lines of fit lying within the angle BOA which also shows that the sign of k must be opposite that of r. As Roos shows (1), k must be some function of the errors. If the errors are assumed equal, $k = 1$ and according to our above analysis the line of fit must be a diameter of the correlation ellipse bisecting chords whose slope equals — 1. This line is shown as line EE' in

(1) *Op. cit.*, p. 12 and 19.

figure 1 and bisects chords parallel to CD. Its equation is found by substituting our values in equation (9). Letting $k = -1$, we then have,

$$y = 0.4745\, x. \tag{20}$$

If we were to change the scale factor as was done above letting $\sigma'_x = 126.4303$ and substituting in equation (9) we would have

$$y = 0.00949\, x' \tag{21}$$

which is exactly what would be obtained by substituting $\frac{x'}{50}$ in equation (20), *i. e.*,

$$y = 0.4745\, \frac{x'}{50} = 0.00949\, x' \tag{22}$$

Again if we let $\sigma_x'' = 0.25286$ and substitute in equation (9) we have

$$y = 4.745\, x'' \tag{23}$$

which is the same as substituting $10\, x''$ in equation (20), for then we get

$$y = 0.4745 \cdot 10\, x'' = 4.745\, x'' \tag{24}$$

It must be remembered that if the scale factor is changed k must change accordingly. For if we let δ_x and δ_y indicate the errors in x and y respectively,

(25) $k = \frac{\delta_y}{\delta_x}$. If one scale is multiplied by a factor β and the primes denote the new values.

$$\delta_x = \beta\, \delta_x' \qquad \text{and } k' = \frac{\delta_y}{\delta_x'} = \beta\, \frac{\delta_y}{\delta_x}, \quad \text{or}$$

$$k' = \beta\, k \tag{26}$$

A very interesting case arises when we have some reason to believe that the error in all the variables are of the same relative order. In that case we may make the assumption that the errors are proportional to their standard deviations, that is,

$$\frac{\delta_x}{\sigma_x} = \frac{\delta_y}{\sigma_y}. \tag{27}$$

In this case k will equal σ_y/σ_x. Substituting this value of k in Eq. (9), remembering that k must have the opposite sign r_{xy}, we have, for positive correlation,

$$y = \frac{\sigma_y}{\sigma_x} x , \tag{28}$$

which is indipendent of r_{xy}. Geometrically this regression line is the diagonal of the rectangle circumscribing the correlation ellipse. This solution can be looked upon in another manner. If the errors were uncorrelated, then y is a function of x alone, and if there were no errors in x and y, the true relation would be given by

$$y = \frac{E(\sigma_y)}{E(\sigma_x)} x , \tag{29}$$

where $E(\sigma_y)$ and $E(\sigma_x)$ are the mathematical expectations or "true" standard deviations of variables. If all the errors are in y we have Eq. (10) where $\sigma_y > E(\sigma_y)$ due to the errors, and $\sigma_x = E(\sigma_x)$. Therefore, we multiply σ_y by r_{xy} to reduce σ_y and get our estimate of $E(\sigma_y)$. The same reasoning applies to Eq. (12) when all errors lie in x. But when both variables are subject to error, we may make the assumption that the ratio σ_y/σ_x is our best estimate of $E(\sigma_y)/E(\sigma_x)$ and the resulting regression will be as given in Eq. (28).

It is now clear that Roos' general equation is independent of the scale factor. It depends on the assumption regarding errors in x and y, and this assumption becomes more and more important as the correlation coefficient becomes small. So we conclude that the " goodness of fit " will depend, in a large measure, on the accuracy of the assumption concerning the errors in the variables.

V. CASTELLANO

Sugli indici relativi di variabilità e sulla concentrazione dei caratteri con segno

SOMMARIO: 1) Indici di variabilità assoluti, ridotti e relativi. Determinazione della distribuzione massimante quando il carattere presenta limiti non ben determinati. Considerazioni che suggeriscono la opportunità dello studio della variabilità e della concentrazione dei caratteri con segno. Scopo del lavoro. — 2) Valori massimi di alcuni indici assoluti di variabilità nella ipotesi della costanza della quantità complessiva del carattere, ma di un campo di variabilità qualunque. — 3) Considerazioni nel caso in cui gli estremi del campo di variabilità sono discordi. — 4) La curva di concentrazione dei caratteri con segno. Estensione del concetto di concentrazione ai caratteri con segno. — 5) Concentrazione degli scarti della distribuzione binomiale. Valore della concentrazione fornito dalla curva binomiale. Avvertenze sull'uso delle distribuzioni teoriche illimitate. — 6) Esempi.

1. — Ostacoli al confronto della variabilità di due seriazioni mediante gli indici assoluti di variabilità sono: la impossibilità di paragonare fra loro quantità non omogenee, come lunghezze pesi, valori, durate, ecc.; e la difficoltà che oppone al confronto il fatto che gli indici assoluti presentano, in generale, campi di variabilità di diversa ampiezza.

Il primo ostacolo si elimina ragguagliando l'indice di variabilità ad una media, entrambi ragguagliando l'indice di variabilità al valore massimo che esso può raggiungere. Si noti che i due ostacoli, hanno carattere tutt'affatto diverso.

Il primo è intrinseco alla natura delle seriazioni che si considerano, si presenta solo quando si vogliono confrontare seriazioni di caratteri non omogenei, esisterebbe anche se non si potessero pensare altre seriazioni dei due caratteri e si elimina, normalmente, avendo riguardo esclusivamente alle distribuzioni in esame.

Il secondo non è connesso alla natura dei due caratteri e si presenta solo quando le due distribuzioni da confrontare non si

considerano ciascuna a sè, ma relativamente agli insiemi di tutte le distribuzioni dei due caratteri che si possono astrattamente pensare come ugualmente possibili. A differenza del primo, questo si presenta anche nel confronto di due distribuzioni di uno stesso carattere, quando si pensino appartenenti ad insiemi diversi, come nel caso di distribuzioni di reddito tassato, quando i limiti di tassazione sono diversi.

L'indice che si ottiene ragguagliando un indice assoluto di variabilità ad una media, generalmente la media aritmetica, può intendersi anch'esso un indice assoluto di variabilità del carattere quando questo sia misurato col metro della media prescelta, anzi credo che questa sia la sola corretta interpretazione di tali indici, che potrebbero denominarsi « indici ridotti » (denominando « intensità ridotte » i rapporti delle quantità in esame ad una loro media, in generale alla media aritmetica), essendo preferibile riservare la denominazione di « indici relativi » alle quantità omogenee e variabili tra 0 e 1 che si ottengono ragguagliando gli indici assoluti al loro massimo (1).

---

(1) Il solo criterio possibile di scelta tra indici assoluti, ridotti e relativi sta nello scopo che ci si propone col confronto della variabilità. Si adotterà, a seconda dei casi, l'indice (ad es.: differenza media assoluta, per cento rispetto ad una media, per cento rispetto al massimo) che ha il significato che interessa. V. le esemplificazioni che ne dà il GINI in *Variabilità e Mutabilità*, (« Studi economico-giuridici pubblicati a cura della Facoltà di Giurisprudenza della R. Università di Cagliari », Bologna, Cuppini, 1912, ristampata nel volume : *Memorie di Metodologia Statistica*, di imminente pubblicazione), alla quale mi pare che non sia nulla da aggiungere. Dovremmo anche menzionare, oltre a questi indici, quelli che si ottengono ragguagliando gli indici assoluti al campo di variabilità del carattere, proposti dallo CZUBER (*Beitrag zur Theorie statistischer Reihen*, « Versicherungswissenschaftlichen Mitteilungen », Neue Folge, B. 9, H. 2, Wien, 1914). Questi indici sono da considerarsi indici relativi, in quanto il campo di variabilità può considerarsi proporzionale, se n non è troppo piccolo, ai massimi degli indici assoluti di variabilità (scostamenti semplice e quadratico medio dalla media aritmetica e dalla mediana, differenze semplice e quadratica media, ecc.), calcolati rispetto a tutto l'insieme delle distribuzioni del carattere, nella sola ipotesi che questo sia limitato. Per tali massimi, che il GINI aveva già considerati in *Variabilità e Mutabilità*, V. la memoria di questo stesso autore : *Sul massimo degli indici di variabilità assoluta e sulle sue applicazioni agli indici di variabilità relativa e al rapporto di concentrazione*, « Metron », Vol. VIII, n. 3, 1930. Io non considero questi indici nel testo, ma per osservazioni sul loro uso V. la ristampa di *Variabilità e Mutabilità*, nel volume : *Memorie di Metodologia statistica*, già cit.

Diciamo *distribuzione massimante* di un carattere la distribuzione di massima disuguaglianza (1).

Il problema che ci si pone è questo: la quantità di carattere che si riferisce ad una determinata distribuzione si può pensare ripartita in infinite maniere diverse tra gli elementi della distribuzione stessa. Per quale ripartizione la disuguaglianza tra i dati diventa massima?

Il male è che non si ha un metro assoluto per la misura della disuguaglianza tra n dati, e si formulano, della disuguaglianza, tanti giudizi quanti sono gli indici di variabilità ai quali si ricorre per la misura (2), e quindi sorge l'inconveniente di dover definire la distribuzione massimante in relazione ai vari indici particolari, e non avere, di conseguenza, un riferimento comune per tutti gli indici. Ma l'inconveniente è più teorico che reale, in quanto la massa degli indici di variabilità più usati (scostamenti medi e differenze medie dei vari ordini) assumono il loro massimo valore per la stessa distribuzione, e quindi inducono ad una unica definizione della distribuzione massimante: *distribuzione di massima disuguaglianza è quella nella quale tutta la quantità del carattere è riunita nel numero minimo di elementi* (3).

Non tutti gli indici, però, concordano in questa definizione, come, ad es., la differenza interquartile e tutti gli indici simili, ad esclusione del campo di variazione, che, per tale distribuzione, assumono un valore nullo, invece che massimo. Ma per quanto riguarda questi indici, il difetto non è davvero della definizione, ma degli indici, che non soddisfano neppure ad una condizione, quale quella di assumere il valore zero solo in corrispondenza della distribuzione uniforme, che parrebbe essenziale per una misura della variabilità.

Il caso della differenza interquartile e degli altri indici di posizione è ovvio e non mi sembra che l'uso di essi sia diffuso,

---

(1) Cfr. C. GINI: *Sul massimo degli indici di variabilità assoluta*, ecc., già cit.

(2) L'argomento è stato da me trattato di proposito in: *Recente letteratura sugli indici di variabilità*, «Metron», Vol. XII, n. 3, 1934-XIII.

(3) Non vedo come si potrebbe stabilire per via diversa una definizione, tutt'altro che intuitiva, come questa. Si pensi che, se l'estremo superiore dal carattere è sufficientemente elevato, in tale distribuzione di «massima disuguaglianza», $n - 1$ quantità sono uguali fra loro.

nè che essi si adoperino alla cieca, senza riguardo alla forma delle distribuzioni relative; ma l'aver potuto definire la « distribuzione massimante » in relazione ad un considerevole complesso di indici di preciso significato costituisce indubbiamente un vantaggio, dal punto di vista della critica degli indici di variabilità.

La definizione data di distribuzione massimante riguarda i caratteri che ammettono un limite inferiore nullo, ma esistono caratteri che, per definizione, non possono presentare valori inferiori ad un dato limite, non nullo, come i patrimoni o i redditi soggetti a tassa, il numero dei vani per abitazione, il numero degli addetti per azienda, ecc., ed altri che non presentano in pratica valori inferiori a certi limiti, ma sono tali che, fissati per essi dei limiti inferiori effettivamente raggiunti, si può sempre presentare qualche caso nel quale il carattere presenti un valore ancora minore, quali, per esempio, la statura e i caratteri antropometrici in generale ; i tempi impiegati dai corridori a coprire un determinato percorso in una gara di corsa, i salari, il costo di produzione, ecc.

Analogamente, non sempre, anzi ancora meno frequentemente, è fissato il limite superiore dei valori che il carattere può raggiungere. Di fronte ad alcuni caratteri limitati per definizione, in genere espressi da rapporti di composizione e di derivazione specifica, esistono caratteri i quali presentano un massimo naturale non solo praticamente, ma anche teoricamente irraggiungibile, come i quozienti di nuzialità e di natalità, che presentano un limite superiore aritmetico 1 ; altri che non presentano nessun massimo naturale, ma che si pensano ordinariamente limitati, come la statura, il peso, e in genere i caratteri antropometrici, la durata della vita, i salari, i prezzi, e così via ; altri, invece, che si considerano praticamente illimitati, pur non essendo tali, come i redditi e i patrimoni. Osserviamo però, a questo proposito, che quando il limite superiore oltrepassa certamente la quantità complessiva di carattere delle distribuzioni in esame, la sua determinazione precisa non ha nessun interesse ai fini della determinazione della distribuzione massimante.

Notiamo che, nella generalità dei casi, si tratta di una indeterminazione dei limiti reale, anzichè di una imperfetta nostra conoscenza di limiti reali precisi. Tale indeterminazione costituisce un ostacolo alla determinazione dell'indice di variabilità relativo.

Questo ostacolo si può superare: *a*) supponendo nullo il limite inferiore indeterminato e infinito il limite superiore indeterminato, e, in questo caso, i risultati ai quali si perviene debbono essere esplicitamente vincolati a questa ipotesi; *b*) definendo opportunamente il carattere in esame in maniera più restrittiva, cioè sostituendo al dato carattere un altro, variabile tra estremi sicuramente determinati; *c*) imponendo al carattere i limiti che suggerisce la curva analitica che ne rappresenta la distribuzione teorica.

Questi metodi vanno naturalmente studiati caso per caso, con molta oculatezza. Il metodo *c*) non dà, in pratica, maggior garanzia di risultati, se non è applicato solo quando esista o si possa ammettere una distribuzione teorica di forma ben determinata e rigorosamente determinata agli estremi.

Quando si abbia riguardo a distribuzioni teoriche illimitate sia a sinistra che a destra, come la curva di Gauss, si presenta particolarmente importante, ai fini della determinazione degli indici di variabilità relativi, la considerazione dei caratteri con segno. Infatti è possibile, mediante translazione, rendere tutte dello stesso segno le quantità di una distribuzione che abbia almeno un limite finito (e la translazione non altera i valori degli indici assoluti di variabilità); ma non è possibile fare altrettanto quando entrambi i limiti sieno infiniti.

A parte questa considerazione, lo studio diretto della variabilità dei caratteri con segno ha un interesse a sè, trascurato, probabilmente, finora, per mancanza di mezzi idonei allo scopo. Ad esempio, il profitto è una variabile che può assumere certamente valori negativi e positivi; qual'è la concentrazione dei profitti nelle varie categorie di industrie? Non mi consta che alcuno abbia cercato di rispondere finora a questa domanda (1).

---

(1) Il SAIBANTE (*I profitti delle Società per azioni e la concentrazione dei capitali industriali*, « Metron », Vol. IV, n. 1, 1926) ha toccato indirettamente la questione, riprendendo, con più adeguati mezzi, il problema della relazione tra ammontare del profitto e ammontare del capitale, già intrapreso dal DOLIA (*I profitti delle Società per azioni italiane dal* 1882 *al* 1903 *in rapporto al problema delle classi medie*, « Studi economico-giuridici della R. Università di Cagliari », Anno III, parte III, 1911). Egli ha misurato la concordanza tra i due fenomeni mediante l'indice di cograduazione e il coefficiente della retta interpolatrice dei vari valori dei profitti, relativi a classi crescenti di capitali. Se la relazione tra profitto e capitale fosse pre-

Scopo del presente lavoro è la estensione delle formule che dànno il valore massimo degli indici assoluti di variabilità più comuni e del concetto di « concentrazione » ai caratteri con segno; di mostrare gli inconvenienti che, all'applicazione del metodo *c*) importa l'adozione di curve teoriche illimitate e di indicare come convenga eliminare tali inconvenienti nel caso della curva di Gauss.

2. — Il Gini ha determinato, nel 1930, nel caso di seriazioni discrete, i valori massimi che possono assumere alcuni indici di variabilità assoluti (scostamenti semplice e quadratico medio dalla media aritmetica e dalla mediana, differenza semplice media con e senza ripetizione) nella ipotesi di caratteri (positivi) limitati solo superiormente (1). Egli ha determinato pure, nel 1931, l'espressione del massimo della differenza media anche quando il carattere ammette un valore minimo diverso da zero (2).

Per gli altri indici si determinano facilmente con analogo procedimento le formule relative al caso di un carattere avente limite inferiore > 0, ma non è neppure necessario ricorrere a tale via in quanto esse si ottengono immediatamente dalla formula determinata dal Gini nella ipotesi di un carattere limitato solo superiormente (con limite inferiore nullo) quando si abbia presente: 1) che la distribuzione massimante (distribuzione che fornisce i massimi valori degli indici di variabilità assoluta, fermo restando il numero degli elementi e l'ammontare totale del carattere della distribuzione data) è la stessa per tutti gli indici; 2) che per la differenza media la distribuzione massimante è quella che compete, nella ipotesi di limite inferiore nullo, alla distribuzione che si ottiene dalla data togliendo a ciascun elemento una quantità di carattere pari al limite inferiore del carattere (che ha quindi la stessa curva di frequenza della data, ma translata verso sinistra, di una quantità uguale al limite inferiore del carattere).

---

cisamente lineare, la concentrazioue dei primi sarebbe precisamente uguale a quella dei secondi, ove si fosse supposto il profitto un carattere non negativo, come il capitale. Poichè tutti i profitti considerati erano in effetto positivi, il risultato sarebbe stato corretto, se esplicitamente riferito alle Società in attivo. Ma la relazione lineare, è, invece, solo largamente approssimativa.

(1) C. GINI: *Sul massimo degli indici di variabilità assoluta*, ecc. già cit.

(2) C. GINI: *Intorno alle curve di concentrazione*, XX^a Sessione dell'Istituto Internazionale di Statistica, Madrid, 1931 e « Metron », Vol. IX, n. 3-4, 1932.

Risulta quindi che tali formule si ricavano da quelle del Gini, sostituendo, al posto delle quantità che compaiono in esse, la differenza tra dette quantità e il limite inferiore. Considerando distribuzioni discrete, il Gini ha dovuto distinguere, per il limite inferiore nullo, il caso in cui la quantità totale del carattere è un multiplo intero del limite superiore, dal caso che non è tale. Nelle distribuzioni continue, che noi qui consideriamo, non è lungo a prendere in esame il secondo caso, in quanto non si considera in esso il numero degli elementi ai quali la distribuzione massimante assegna la quantità di carattere massimo possibile, bensì la frazione che questi costituiscono sul totale degli elementi della distribuzione.

Indicato con n il numero degli elementi della distribuzione, con A la media aritmetica, con M la mediana, con l e L rispettivamente i limiti inferiore e superiore del carattere, i valori massimi dei suddetti indici di variabilità assoluta sono espressi dalle seguenti formule :

$$\text{Max}\,{}^1S_A = \frac{2(A-l)\,(L-A)}{L-l} \qquad \text{Max}\,{}^1S_M = \begin{cases} A-l \\ \text{se } L \gtrless 2A \\ L-A \end{cases}$$

$$\text{Max}\,{}^2S_A = \sqrt{(A-l)\,(L-A)} \qquad \text{Max}\,{}^2S_M = \begin{cases} \sqrt{(A-l)\,(L-l)} \\ \text{se } L \gtrless 2A \\ \sqrt{(L-A)\,(L-l)} \end{cases} \qquad (1)$$

$$\text{Max}\,\Delta = 2\,\frac{n}{n-1}\,\frac{(A-l)\,(L-A)}{L-l} \qquad \text{Max}\,\Delta_R = \frac{2(A-l)\,(L-A)}{L-l}$$

É appena utile ricordare che queste formule valgono nel caso in cui si attribuisca lo stesso peso a tutti i termini della seriazione e per $L \leq n\,A - (n-1)\,l$. Il caso $L > nA - (n-1)\,l$ equivale al caso di un carattere non limitato nel quale la quantità di carattere eccedente quella minima compatibile col limite inferiore l dei valori del carattere può concentrarsi in un solo elemento, e quindi le espressioni dei massimi considerati assumono il valore delle (1) per $L = n\,A - (n-1)\,l$. È facile vedere che $\frac{A-l}{L-l}$ rappresenta, nel nostro caso, la frazione degli elementi della data distribuzione ai quali tutti la distribuzione massimante attribuisce la massima quantità di carattere L.

Poichè nessuna ipotesi sul segno di l è stato necessario introdurre per la deduzione delle (1), esse sono valide anche per $l < 0$; esse sono valide anche comunque si translati la curva di frequenza della distribuzione data, solidalmente al campo di variabilità del carattere, poichè la distribuzione massimante della data subisce la medesima translazione: quindi le (1) valgono anche per le distribuzioni di caratteri negativi, o positivi e negativi insieme.

È noto che gli indici di variabilità assoluti sono invarianti per translazione delle curve di frequenza delle distribuzioni cui si riferiscono, perchè dipendono solo dalla forma delle rispettive curve di frequenza e non dalla intensità del carattere. Similmente, *gli indici di variabilità relativi, dedotti dal rapporto degli indici assoluti di variabilità al proprio massimo, sono invarianti per translazione della curva di frequenza delle distribuzioni cui si riferiscono, solidali col campo di variabilità del carattere: essi dipendono dalla forma della curva di frequenza e dalla distanza di questa dal polo* (1), *non dalla intensità del carattere.*

3. — Abbiamo illustrato, al numero precedente, i limiti di applicazione delle (1) quando si abbia riguardo ai soli caratteri positivi. Se si considerano i caratteri con segno, occorre aggiungere alla condizione ivi posta ($L \leq nA - (n-1)l$) solo quella di l finito. Quest'ultima condizione è ovvia, in quanto la translazione con la quale è stata ottenuta la precedente estensione, non può essere che finita.

Osserviamo anche che, poichè le frazioni del complesso degli elementi della distribuzione data, che, nella distribuzione massimante, presentano le intensità del carattere minima e massima (che sono rispettivamente $\frac{L-A}{L-l}$ e $\frac{A-l}{L-l}$), dipendono esclusivamente da l e L (A restando costante), la distribuzione massimante non conserva, di regola, anche l'ammontare complessivo delle quantità negative e positive, quando l e L sono di segno discorde. Anzi, in questo caso, la distribuzione massimante consta sia di quantità positive

---

(1) Uso la denominazione di «polo» nel significato di limite inferiore del carattere, da me definito in *Sulle relazioni fra curve di frequenza e curve di concentrazione, e sui rapporti di concentrazione corrispondenti a determinate distribuzioni*. («Metron», Vol. X, n. 4, 1933), anzichè in quello di «origine delle coordinate», che non ha, nella teoria della concentrazione dei caratteri, nessun interesse.

che di quantità negative, anche se la distribuzione data consta di sole quantità positive o negative.

Dal confronto delle quantità di carattere di un dato segno spettanti alla distribuzione data e a quella massimante, risulta facilmente che a quest'ultima distribuzione competono quantità di carattere maggiori, in valore assoluto, che non all'altra (1).

Se si impone la condizione della costanza delle quantità di carattere negativa e positiva della distribuzione data, si ottiene una distribuzione massimante meno disuguale, dal punto di vista degli indici da noi considerati, della precedente, e che necessariamente non gode, come quella, della proprietà di essere invariante per translazione solidale della curva di frequenza della distribuzione che si considera, e del campo di variabilità del carattere. Ciò significa che gli indici di variabilità relativi corrispondenti dipendono, in questo caso, anche dalla intensità del carattere.

La determinazione di tali massimi è stata fatta recentemente, per caratteri non limitati, dal prof. M. de Vergottini (2).

4. — In un precedente lavoro (3) ho già mostrato come si deformi la curva di concentrazione per una translazione della corrispondente curva di frequenza. Ma, partendo allora da una distribuzione di polo zero e di distanza polare nulla, considerai translazioni verso destra, la mia attenzione essendo allora limitata ai soli caratteri positivi. Vediamo ora quale forma assuma la curva di concentrazione quando alcuni, o tutti gli elementi di una distribuzione assumano valori negativi.

Costruiamo la curva di concentrazione C_{f_0} relativa alla distribuzione $y = f(x)$, definita nel campo $0 \vdash c$, portando sulle

---

(1) Infatti, dalla disuguaglianza $\frac{L-A}{L-l} nl \overset{\circ}{\neq} n_1 A_1$, i due membri della quale rappresentano, ordinatamente, la quantità di carattere negativa spettante alla distribuzione massimante e a quella data, si ricava facilmente l'altra: $1 \overset{\circ}{\neq} \frac{n_2}{n}\frac{A_2}{L} + \frac{n_1}{A}\frac{A_1}{l}$, nella quale il primo membro è ovviamente maggiore del secondo. Sicchè il segno $\overset{\circ}{\neq}$ della disuguaglianza di partenza equivale a $\angle$.

(2) M. DE VERGOTTINI: *Relazioni tra gli indici di variabilità dei fenomeni collettivi composti e quelli dei fenomeni collettivi elementari*, Tipografia Failli, Roma, 1936-XIV.

(3) Cfr. V. CASTELLANO: *Sulle relazioni fra curve di frequenza*, ecc., già cit.

ordinate, in corrispondenza a ogni valore di $0 \leq p \leq 1$, l'ammontare del carattere posseduto complessivamente dalla frazione p degli elementi meno forniti del carattere. Fissato un valore di $p = \bar{p}$ la corrispondente ordinata $\bar{q}$ della curva di concentrazione relativa alla distribuzione $y = f(x + \varepsilon)$ (che si ottiene translando di una quantità ε verso sinistra la $f(x)$, si ottiene, sottraendo alla ordinata della C_{f_0} la quantità $\varepsilon\, n\, \bar{p}$, in quanto la stessa frazione $\bar{p}$ degli elementi della f possiede, nella f_ε una quantità di carattere $\varepsilon\, n\, \bar{p}$ meno che nella f_0, e quindi $\bar{q}_{(\bar{p})} = q_{(\bar{p})} - n\,\varepsilon\,\bar{p} : n\,\varepsilon\,\bar{p}$ è anche la differenza tra le ordinate dei punti delle rette di equidistribuzione della f_0 e della f_ε, aventi la stessa ascissa $\bar{p}$. Se immaginiamo l'area di concentrazione della f_0 realizzata mediante un gran numero di cilindri sottilissimi paralleli all'asse delle y e scorrevoli lungo i propri assi, vincolati ciascuno all'estremo superiore ad un anello scorrevole sulla retta di equidistribuzione e i cui estremi inferiori descrivano la C_{f_0}, la C_{f_ε} risulta descritta dagli stessi estremi, quando si faccia ruotare la retta di equidistribuzione della f_0, fino a raggiungere la posizione della retta di equidistribuzione della f_ε.

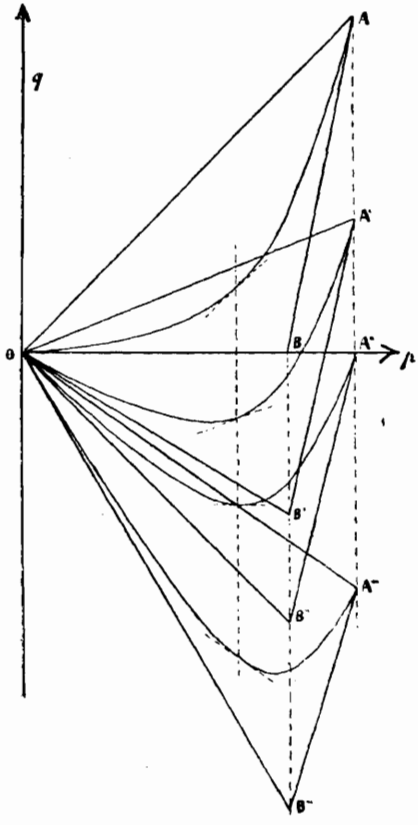


Fig. 1

Finchè $\varepsilon < c$, talchè lo zero risulta interno al campo di definizione $-\varepsilon \dashv c - \varepsilon$ della $f(x + \varepsilon)$, la C_{f_ε} si compone di due rami, il 1° decrescente, dal punto (0,0) al punto $(p(\varepsilon), q(\varepsilon))$; il 2° crescente, da questo al punto $(1, m - n\,\varepsilon)$, essendo m la quantità di carattere relativa alla distribuzione f_0. Il ramo crescente taglia l'asse delle p finchè $\varepsilon \leq A$; per $\varepsilon = A$, A essendo la media aritmetica della f_0 (talchè la media aritmetica A_ε della f_ε risulta nulla), esso incontra detto asse nel punto estremo (1,0); per $\varepsilon > A$,

talchè $A_\varepsilon < 0$, la $C_{f\varepsilon}$ si svolge tutta al di sotto dell'asse $q = 0$; per $\varepsilon \geq c$ essa si riduce al solo 1° ramo.

Come la C_{f0}, la $C_{f\varepsilon}$ volge la sua concavità verso l'alto. Nel punto $p\,(A_\varepsilon) = p\,(A)$ la sua tangente è parallela alla retta di equidistribuzione. Se la f è simmetrica, la $C_{f\varepsilon}$ è simmetrica rispetto alla retta $p = \frac{1}{2}$: se è $A_\varepsilon = 0$ la simmetria è normale, se $A_\varepsilon \neq 0$, la simmetria è obliqua.

La trasformazione che porta la C_{f0} in $C_{f\varepsilon}$ porta anche le tangenti della prima in quelle della seconda. Poichè essa fa scorrere su sè stessa le rette $p = \mathrm{cost.}$, il vertice del triangolo formato dalle tangenti esterne alla $C_{f\varepsilon}$ colla relativa retta di equidistribuzione, si trova sulla stessa retta $p = \frac{c - A}{c}$, sulla quale si trova il vertice della distribuzione di massima concentrazione relativo alla f_0 (supposto $l = 0$, $L = c$).

Dal punto di vista formale, la curva di concentrazione resta definita dalle normali equazioni,

$$p = \frac{1}{n}\int_l^z f(z)\,dz\,, \qquad q = \frac{1}{m}\int_l^z z f(z)\,dz \qquad (2)$$

per $m \neq 0$. Nel caso di $m = 0$, si potrà dividere l'integrale a secondo membro della seconda equazione per

$$m' = \int_A^L z f(z)\,dz = \int_l^A |z|\, f(z)\,dz\,,$$

in maniera da tenere costante ($= 1$) l'ordinata minima della curva; ma ciò non è strettamente necessario.

Anche l'area di concentrazione resta espressa dalla formula

$$S = \frac{1}{2}\left\{\int p\,dq - \int q\,dp\right\}.$$

Ragguagliando il valore di S, relativo alla distribuzione in esame, a quello relativo alla corrispondente distribuzione massimante, si ha il valore di $0 \leq R \leq 1$ (1).

---

(1) Al Wold (*A study on the mean difference, concentration curves and concentration ratio*, « Metron », Vol. XII, n. 2, 1935) che pure ha considerato il caso di variabili che possono assumere anche valori negativi, è sfuggita la possibilità di estendere la validità delle (2) anche ai valori di $z < 0$ mantenendo R nei limiti 0 e 1. Gli è che l'A. ha considerato l'estensione

Dal punto di vista concettuale la estensione ai caratteri con segno non presenta difficoltà. Infatti, continuiamo a dire col Gini (2) che il carattere è tanto più concentrato, quanto maggiori sono le disuguaglianze $p > q$, essendo p e q definiti dalle (1); che tanto più forti in valore assoluto sono le disuguaglianze $p > q$ quanto maggiore è la differenza $p - q$, tanto più forte in valore relativo quanto maggiore è il rapporto

$$R_{(p)} = \frac{p-q}{\frac{A-l}{A}p} \quad \text{oppure} \quad \frac{p-q}{(1-p)\,\frac{L-A}{A}},$$

a seconda che è $p \lessgtr \frac{L-A}{L-l}$. $R\,(p)$ rappresenta il coefficiente per cui bisogna moltiplicare la massima differenza possibile tra p e q, compatibile con i limiti l e L e la costanza di A, per ottenere la differenza effettiva $p - q$. $R\,(p)$ varia tra 0 (caso di equidistribuzione e 1 (caso di massima concentrazione).

Come misura della concentrazione, si assuma una media ponderata degli $R\,(p)$, assumendo come pesi le quantità $\frac{A-l}{A}p$ e $(1-p)\frac{L-A}{A}$, a seconda che $p \lessgtr \frac{L-A}{L-l}$.

Osservato che

$$\int (p-q)\,dp = \frac{\Delta_R}{4A}\,;\quad \frac{A-l}{A}\int_0^{\frac{L-A}{L-l}} p\,dp + \frac{L-A}{A}\int_{\frac{L-A}{L-l}}^{1} p\,dp = \frac{(A-l)(L-A)}{2A(L-l)},$$

risulta:

$$R = \frac{(L-l)}{2(L-A)(A-l)}\Delta_R,$$

che è la formula alla quale si perviene per R mediante le (1).

---

solo da un punto di vista astratto, ed è partito dalla definizione di $R = 2S$, che non deve considerarsi valida quando il limite inferiore dei valori del carattere è diverso da zero. Già il Gini (*Sul massimo degli indici di variabilità*, ecc., già cit.), però, aveva introdotto nella espressione del rapporto di concentrazione un coefficiente di correzione nel caso di $l > 0$.

(1) Cfr. C. Gini: *Sulla misura della concentrazione e della variabilità dei caratteri*, « Atti del R. Istituto Veneto di Scienze, Lettere ed Arti », Anno 1913-14, Tomo LXXIII, P. seconda, pagg. 4-5 dell'estratto.

Non è forse superfluo osservare che la definizione non cade in difetto per $m = o$, solo che si sostituisca alla seconda equazione delle (2), la

$$q = \frac{1}{m'} \int_l^x z f(z)\, dz \text{ oppure } q = \int_l^x z f(z)\, dz .$$

Gli R_p si ottengono ugualmente uguagliando le differenze $p - q$ ai loro massimi, e R come media ponderata degli R_p.

5. — Consideriamo, a mo' d'esempio, la distribuzione binomiale nel caso $p = q$. Nello schema di Bernoulli, $\binom{n}{np + l} p^{np+l} q^{nq-l}$ è la frequenza teorica di uno scarto l tra il numero delle palline bianche estratte in n successive estrazioni da un'urna contenente palline bianche e nere nella proporzione p/q, quando si consideri un numero sufficientemente grande di serie di estrazioni.

La curva di concentrazione consta dei segmenti che congiungono ordinatamente i punti: (0,0), $(q^n, -npq^n)$, $(q^n + npq^{n-1}, -npq^n - [np - 1]pq^{n-1})$, ecc. (Nella Fig. 2 essa è disegnata nella ipotesi $p = q = \frac{1}{2}$ e $n = 10$).

Per $p = q = \frac{1}{2}$ è noto che

$$\Delta_R = \frac{(2n - 1)!}{2^{2n-1}[(n-1)!]^2} = \sqrt{\frac{n}{\pi}} \quad (1)$$

quando n è sufficientemente grande. Ma già per $n = 8$ l'errore che si commette sostituendo alla prima la seconda espressione è $< 2\ \%$.

---

(1) La formula fu data, in *Variabilità e Mutabilità*, dal Gini che la determinò empiricamente. Essa fu dimostrata quindi dal Cantelli (*Sulla differenza media con ripetizione*, « Giornale degli Economisti e Rivista di Statistica », febbraio, 1913). Il valore $\sqrt{\frac{n}{\pi}}$ è stato ricavato anche dal Gini, trasformando la espressione esatta con la formula di Stirling, supposto n grandissimo ed è quella che compete alla curva normale. Infatti per la curva normale si ha:

$$\Delta/\sigma\sqrt{2} = \frac{2}{\pi} \int_{-\infty}^{\infty} e^{-z^2} \int_{-\infty}^{x} (x - z)\, e^{-z^2}\, dz dx = \sqrt{\frac{2}{\pi}} \text{ e quindi}$$

$\Delta = \frac{2\sigma}{\sqrt{\pi}} = 2\sqrt{\frac{npq}{\pi}}$. Da questa, per $p = q = \frac{1}{2}$, si ha $\Delta = \sqrt{\frac{n}{\pi}}$.

Poichè in questo caso $A = 0$, $l = \frac{n}{2}$, $L = \frac{n}{2}$, è Max $\Delta_R = \frac{n}{2}$, e quindi $R = \frac{(2n)!}{2^{2n-1}(n!)^2} = \frac{2}{\sqrt{n\pi}}$ (3)

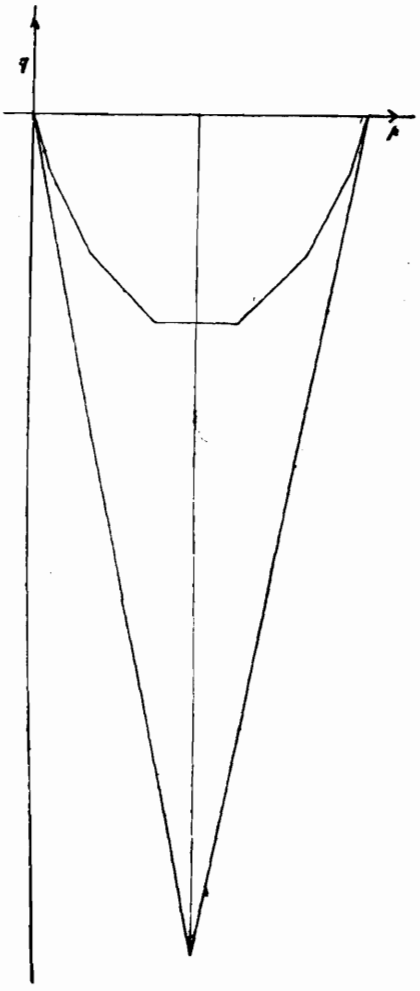


Fig. 2

Tenuto conto delle unità di misura adottate l'area di concentrazione S risulta:

$$S = \frac{\Delta_R}{4} = \frac{1}{4}\sqrt{\frac{n}{\pi}} .$$

La ordinata minima della curva, che è uguale alla metà dello scostamento semplice medio relativo quando si assuma come unità di misura l'altezza del triangolo di massima concentrazione, risulta

$$\min h \simeq \frac{1}{2}\sqrt{\frac{n}{2\pi}} .$$

Se facciamo tendere n all'infinito, il diagramma binomiale si approssima indefinitivamente alla curva binomiale. La (3) fornisce per $n = \infty$, $R = 0$. E appare utile osservare che si sarebbe pervenuti allo stesso risultato anche considerando gli scarti relativo e ridotto. L'area di concentrazione relativa alla distribuzione degli scarti assoluti cresce proporzionalmente a $\sqrt{n}$, mentre il corrispondente triangolo di massima concentrazione cresce proporzionalmente a n: l'area di concentrazione relativa alla distribuzione degli scarti ridotti è costante, mentre il corrispondente massimo cresce come $\sqrt{n}$: l'area di concentrazione relativa alla distribuzione degli scarti relativi decresce come $\frac{1}{\sqrt{n}}$, mentre il corrispondente massimo resta costante.

Ma la curva binomiale non ha interesse come legge limite della distribuzione binomiale, bensì come legge approssimata di questa, per n finito. Sarebbe errato, in questo caso, affermare che la concentrazione è nulla perchè la sua area di concentrazione è finita, mentre il corrispondente massimo è apparentemente infinito, essendo la curva illimitata sia a sinistra che a destra. L'area del triangolo di massima concentrazione, data dalla curva approssimata, è infinita, mentre quella del triangolo relativo alla distribuzione esatta, è finita. La sostituzione non è quindi lecita, in questo caso, come è lecita invece la sostituzione della spezzata di concentrazione relativa al diagramma binomiale con la curva di concentrazione fornita della curva binomiale, perchè questa costituisce, per n non troppo piccolo, una buona approssimazione di quella.

In tale caso quindi come valore del rapporto di concentrazione deve assumersi quello dato dalla (3), e non quello dato dalla curva binomiale.

La curva binomiale si definisce anche indipendentemente dalla distribuzione binomiale, come legge normale della probabilità. La variabile casuale, definita dalla legge normale, ha effettivamente concentrazione nulla e, in generale, variabilità relativa nulla.

Questo risultato non osta al fatto che la distribuzione definita dalla legge normale non è la distribuzione uniforme, perchè essa è una distribuzione teorica, che verte su un numero infinito di casi e su quantità illimitate di carattere, e non è strano che proprietà valide incondizionatamente al finito, non sieno valide quando si allarghi indefinitamente il campo nel quale si considerano.

Appunto perchè suppone un carattere illimitato nei due sensi e un numero infinito di osservazioni, la curva binomiale, nella sua accezione di « legge normale di probabilità » come in quella di « legge degli errori accidentali », al pari di ogni altra curva di distribuzione asintotica che pretenda il nome di legge, dovrebbe essere riguardata con alquanto sospetto dagli statistici, o almeno adoperata con estrema prudenza.

Le distribuzioni statistiche constano di quantità finite in numero limitato. È possibile verificare l'adattamento ai dati di una parte solamente, ben limitata, della distribuzione teorica. Nessun adattamento, per quanto buono, autorizza ad ammet-

tere che per un numero infinito di osservazioni la legge di distribuzione del fenomeno sarebbe asintotica, quand'anche una legge asintotica non fosse, a priori, del tutto assurda. D'ordinario questa estrapolazione è assolutamente illecita, e può portare a risultati aberranti, non solo nel campo teorico, ma anche in quello pratico, come nel caso della determinazione degli indici di variabilità relativi.

Credo quindi che sia conveniente attribuire un campo di variabilità limitato alla curva normale, in tutte le sue accezioni. Se essa si adopera come legge teorica di distribuzione di una serie di frequenze (ad es. rapporto di mascolinità delle nascite), e ciascuna frequenza è determinata da un numero n di prove, allora i limiti del carattere sono $-p$ e $+q$ (1) e R è dato dalla (3); ma se la curva di Gauss si adopera come legge di distribuzione di un carattere cui non si possa dare il significato di una frequenza e che non ha dei limiti ben certi, determinabili a priori, come ad es. la statura, allora occorrerà esaminare la questione caso per caso. Per analogia col caso precedente, si potrebbe pensare di assumere ancora la (3) come espressione di R, determinando l'n che compare nella (3) uguagliando lo scarto teorico $\frac{\sqrt{n}}{2}$ che si otterrebbe, per $p = q = \frac{1}{2}$, se si trattasse di una curva di frequenza assoluta, allo scarto empirico $\sigma = \sqrt{\frac{\Sigma (x_i - A)^2}{s}}$, ma occorre controllare che per il carattere in questione i limiti che ne derivano, $l = -2\sigma^2$, $L = 2\sigma^2$, non risultino assurdi.

Per la curva degli errori accidentali credo convenga senza altro assumere la (3) come valore teorico di R, invece di zero.

6). — *a*) Consideriamo anzitutto una distribuzione teoricamente del tipo binomiale.

Ecco i risultati di 37 serie di 100 estrazioni da un'urna contenente un congruo numero di palle bianche e palle nere, fatte eseguire dallo Czuber ai suoi allievi (2):

---

(1) Naturalmente si assumerà per p la frequenza media del carattere che si desume dalle s n osservazioni, se s è il numero delle frequenze considerate.

(2) G. CASTELNUOVO: *Calcolo delle probabilità*, Zanichelli, Bologna, 1933, vol. I, pagg. 118-9.

valore dello scarto

$-8 \quad -7 \quad -6 \quad -5 \quad -4 \quad -3 \quad -2 \quad -1 \quad -0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 6 \quad 8$

numero delle serie (4)

$1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 4 \quad 1 \quad 4 \quad 4 \quad 2 \quad 4 \quad 1 \quad 2 \quad 5 \quad 2$

I valori teorici di σ e di Δ_R sono rispettivamente:

$$\sigma = \sqrt{\frac{n}{4}} = 5 \qquad \Delta_R = \sqrt{\frac{n}{\pi}} = 5{,}64$$

mentre i corrispondenti valori empirici sono:

$$\sigma' = 4{,}28 \qquad \Delta'_R = 5{,}01.$$

I limiti effettivi degli scarti dal valore medio sono $-l = L = 50$. Se non conoscessimo il numero delle prove di ciascuna serie, calcoleremmo, in base al valore di σ': $\frac{n'}{2} = 2\,\sigma'^2 = 36{,}64$ e, in base al valore Δ'_R: $\frac{n}{2} = \frac{1}{2}\pi\,\Delta'^2_R = 39{,}43$

Il valore teorico di R (che si verificherebbe in un numero infinito di serie) è

$$R = \frac{1}{5\sqrt{\pi}} = 0{,}113\,;$$

mentre il valore effettivo è:

$$R' = \frac{2\,\Delta'_R}{100} = 0{,}100.$$

Se non conoscessimo il numero n delle prove di ciascuna serie, daremmo per R e R' la stessa valutazione $R = \frac{1}{\sigma'\sqrt{\pi}} = 0{,}132$ in base al valore di σ' e $R = \frac{2}{\pi\,\Delta'_R} = 0{,}127$ in base al valore di Δ'.

Questi valori sono entrambi superiori al valore teorico di R, in accordo colla teoria (v. es. *c*), essendo σ' e Δ' entrambi risultati minori dei corrispondenti σ e Δ.

Come si vede risulta per R e R' un valore molto piccolo, ma non nullo; la sola considerazione della (4) non avrebbe permesso la determinazione di R, data la grande difficoltà della valutazione dei limiti l e L.

b) Consideriamo la seguente distribuzione delle somme assicurate nella forma mista, per classi quinquennali di età degli assicurati, distribuzione interpolata dall'Elderton colla curva normale (1).

Anno centrale di ciascuna classe di età	17	22	27	32	37	42	47	52	57	62	67
Somme assicurate (in migliaia di sterline)	18	484	124	213	281	295	185	104	40	15	3

Lo scarto quadratico medio della distribuzione è $\sigma = 8{,}87$. Questo valore di σ fornisce, per la differenza media, il valore $\Delta_R = 9{,}98$ mentre il calcolo diretto fornisce $\Delta'_R = 9{,}85$.

Non soddisfacente è in questo caso, il valore di n che fornisce la curva normale, nella ipotesi di una distribuzione binomiale. Infatti σ fornisce: $\frac{n}{2} = 2\,\sigma^2 = 157$ e Δ_R fornisce parimenti $\frac{n}{2} = \frac{1}{2}\pi\,\Delta^2_R = 157$, valori entrambi che forniscono limiti assurdi per la distribuzione del carattere, ai quali corrisponderebbero i valori di $R = 0{,}064$ in base a σ e $R = 0{,}063$ in base a Δ'_R.

Assumendo, in questo caso, $l = 0$, $L = 2\,A = 78{,}40$ risulta $R = \frac{\Delta_R}{A} = 0{,}251$, valore questo che deve ritenersi molto più plausibile del precedente.

c) Consideriamo la curva degli errori accidentali.

Supposti gli errori illimitati, questa fornisce, come abbiamo detto, $R = 0$.

Supponendo la curva limitata, come la curva binomiale, si ha: $R = \frac{1}{\sigma\sqrt{\pi}} = h\sqrt{\frac{2}{\pi}}$, cioè la concentrazione cresce colla *precisione* della distribuzione.

Determiniamo ora la concentrazione dei valori assoluti degli errori accidentali.

Nella ipotesi $L = \infty$, il valore di R per questa distribuzione si ottiene dalla formula: $\frac{\Delta_R - A}{A}$ dove Δ_R ha lo stesso valore

---

(1) W. P. Elderton: *Frequency curves and correlation*, London, 1906.

che per l'intera distribuzione, essendo questa simmetrica, e, $A = \sigma\sqrt{\frac{2}{\pi}}$. Si ha quindi $R = \sqrt{2} - 1$ (1).

Se il carattere si considera limitato superiormente, con: $L = \frac{\sigma^2}{2}$, il precedente valore va diviso per $\frac{L - A}{L} = 1 - \frac{2\sqrt{2}}{\sigma\sqrt{\pi}}$ e quindi $R = \frac{\sqrt{2} - 1}{1 - \frac{2\sqrt{2}}{\sigma\sqrt{\pi}}} = \frac{\sqrt{2} - 1}{1 - \frac{4h}{\sqrt{\pi}}}$.

(1) Cfr. E. J. GUMBEL: *Das Konzentrationsmass*, « Allgemeines Statistisches Archiv », B. 18, H. 2, 1928 e V. CASTELLANO: *Sulle relazioni*, ecc., già cit.

PER OTTESTAD

The exponential frequency function and frequency distributions

Introduction

The fundamental problem in mathematical statistics is that of finding the mathematical function which can be used as an approximate expression of the variation of a number of given observations. This is a problem of finding the best function of equalization. Another problem is that of finding criteria by means of which it can be proved whether or not the variation of given observations can be replaced by a given theoretical function.

In biological investigations one is very often faced with frequency distributions which are arranged from measurement data, measurement of body length or body weight, length or weight of a certain organ, or with distributions which are arranged from data of the number of organs in the organisms. To find a mathematical expression for such distributions has been an important problem for a very long time. Quetelet who introduced the exact measurement methods in the Anthropometry, found that a great number of such distributions could be repalced by the frequency function which was introduced by Gauss. This function which is now known as the exponential or typical frequency function, is given by the formula :

$$f(x) = \frac{1}{\sqrt{2\,\pi\, m_2}}\, e^{-\frac{x^2}{2\,m_2}} \qquad (1)$$

wherein m_2 is the second middle moment.

The investigations of Quetelet pointed out that the exponential frequency function was a very good interpolation formula, but for a long time this function was taken as a natural law. Un-

questioning faith dominated over a long period and many investigators could not lay off the restraint which the exponential function put upon the work. It was very often taken as a fact that frequency distributions arranged from measurement data had to take the form of this function, and very seldom the actual form of the distribution was investigated.

In my opinion the problem should now be taken up again and discussed upon a new basis. In our time there is a tendency among scientists to suppose that any mathematical function is a priori as valuable as every other function. I clearly see that this is a valuable standpoint as far as the functions have only to be used for interpolation purposes. But it should not be forgotten that in this connection the exponential frequency function has some advantages over all other mathematical functions. It is proved that the distribution of a quantity which is the sum of a number of quantities, tends to take the form of the exponential frequency function. Most biological quantities may be taken as such a sum or at any rate as due to a great number of elementary causes. Whittaker and Robinson (*) say « Thus the deviation of the actual chest measure of an individual from the average may be regarded as the sum of a very great number of very small deviations (positive or negative) due to the separate factors in the heredity and environment of the individual ». In my opinion this philosophical hypothesis gives an excellent explanation of why frequency distributions can be so often replaced by the exponential frequency function.

More fundamentally than this theoretical argumentation is, however, the evidence of experience. Skewness and irregularities have, in my opinion, been ascribed a too great significance. Frequency distributions which are arranged from a material of measurement data which has not been carefully treated beforehand, very often show such irregularities. But very often these irregularities are only caused by lack in homogenity of the material. If, therefore, the material is arranged in groups according to some sign, the new distributions are as a rule more regular and very often exponential.

I think that it will be a very fruitful work to investigate

(*) E. T. WHITTAKER and G. ROBINSON. *The Calculus of Observations*. London, 1924.

the frequency distribution of measurement data along new lines. Experience will certainly show that skewness and irregularities with which the biologists are so often faced, depend upon the lack in homogenity of the material, and that these irregularities can be removed by a grouping of the material. The scientists have not been fully aware of the effect which the lack in homogenity has upon the distribution form.

The aim for this paper is not to explain the agreement between frequency distributions and the exponential frequency function, but to introduce a new method by means of which we can investigate whether or not a distribution has exponential form. First I shall deal with the exponential function of one variable and later with the same function of two variables.

THE EXPONENTIAL FREQUENCY FUNCTION OF ONE VARIABLE.

In the course of time there have been introduced several methods for investigating whether a distribution has a form given by a certain mathematical function. Here is not the place to give a description of all these methods. I shall only deal shortly with some methods which have been used for the purpose of investigating if the distribution is exponential. Thiele (*) has proved that the semi-invariants of the exponential frequency function are equal to zero with the exception of that of the second order (the square of the standard deviation). Steffensen (**) supposes that this can be used as a criterion that the distribution is exponential. Steffensen prefer, however, as far as I have seen, another method. He calculates the theoretical frequency values and compares these with those observed. In most cases the observed frequencies give the number of individuals falling within a number of class intervals. If $h\,(x)$ is the number of individuals falling within the class interval $(x_a - x_b)$, we have to compare this number with the theoretical number found by the formula :

$$h'\,(x) = N \int_{x_a}^{x_b} f\,(x)\,d\,x$$

(*) THIELE. *Theory of Observations*. London, 1903.

(**) J. F. STEFFENSEN. *Matematisk Iagttagelseslære*. Kobenhavn, 1923.

We have to calculate such integrals for all class intervals. Of course these integrals can be calculated, but the calculations entail much difficult work. Some investigators satisfy themselves by inserting in the function the value of the variable laying in the middle of the class intervals. The values found in this way are in some cases nearly correct, but in most cases they deviate very much from the correct values. Later I shall show a more convenient method by means of which the theoretical frequency values can be calculated.

Guldberg (*) who has worked with some discontinuous functions, has treated the problem in a different way. From the difference equation of the functions he has succeded in deriving criteria by means of which it is very easy to prove whether a given distribution can be replaced by a certain discontinuous function ; the binomial, the Poisson, the Pascal and the hypergeometrical frequency function. The binomial function

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

has the difference equation

$$f(x+1) = \frac{p}{1-p} \cdot \frac{n-x}{x+1} f(x)$$

From this equation Guldberg has deduced the function

$$\psi(x) = \frac{f(x+1)}{f(x)}(x+1) + \frac{m_1 - m_2}{m_2} x = \frac{m_1^2}{m_2}$$

where m_1 and m_2 are the first and the second middle moment. If an observed distribution satisfies this equation or if

$$\alpha(x) = \frac{m_2}{m_1^2} \cdot \psi(x)$$

with all values of the variable is equal to unity and $m_1 > m_2$, the distribution can be replaced by the binomial frequency function.

Guldberg calls this $\alpha(x)$ the approximation coefficient. This coefficient does not only give a total description of the agreement or disagreement between the distribution and the binomial function,

---

(*) ALF GULDBERG, *On discontinuous frequency functions and statistical series.* « Skandinavisk Aktuarietidskrift », Uppsala, 1931.

but it also gives a description of the agreement or disagreement for all values of the variable. In practice it is not only important to see if the observed distribution agrees or disagrees with the function, but also to see if there is agreement over the whole range of variation.

If n in the binomial function increases to infinity, the function will take the form of the exponential frequency function. Guldberg's approximation coefficient of the binomial function can, therefore, be used when the problem is to find if a distribution can be replaced by the exponential function. This method has been used by the author in an earlier paper (*). This method has that disadvantage that it is based upon the assumption that the zero of the variable is known. The approximation coefficient is given by:

$$\alpha(x) = \frac{m_2}{m_1^2}\,\psi(x) = \frac{f(x+1)}{f(x)}\,(x+1)\,\frac{m_2}{m_1^2} + \frac{m_1 - m_2}{m_1^2}\,x$$

If now x is substituted by $(x + c)$, we find:

$$\alpha(x+c) = \frac{f(x+1)}{f(x)}\,(x+c+1)\,\frac{m_2}{(m_1+c)^2} + \frac{m_1 + c + m_2}{(m_1+c)^2}\,(x+c)$$

The problem is now to find those values of c which make $\alpha(x + c) = 1$.

We find:

$$\alpha(x+c) = \frac{1}{(m_1+c)^2}\left[\left(\frac{f(x+1)}{f(x)}\,(x+1) + \frac{m_1 - m_2}{m_2}\right)\frac{m_2}{m_1^2}\cdot m_1^2 + \right.$$

$$\left. + \frac{f(x+1)}{f(x)}\,c\,m_2 + c\,(x + m_1 + c - m_2)\right]$$

$$= \frac{1}{(m_1+c)^2}\left[\frac{f(x+1)}{f(x)}\,c\,m_2 + c\,(x + m_1 + c - m_2) + m_1^2\right]$$

We can now try to find those values of c which will make $\alpha(x + c) = 1$.

We set, therefore, $\alpha(x + c) = 1$, i. e.

$$\frac{1}{(m_1+c)^2}\left[\frac{f(x+1)}{f(x)}\,c\,m_2 + c\,(x + m_1 + c - m_2) + m_1^2\right] = 1$$

(*) PER OTTESTAD, *A mathematical method for the study of growth.* «Hvalrådets Skr. Nr. 7. Det norske Videnskaps-Akademi i Oslo», Oslo, 1933.

From this we find:

$$\left[\frac{f(x+1)}{f(x)}\, m_2 + x - m_1 - m_2\right] c = c \cdot \varphi(x) = 0$$

From the difference equation will be found that

$$\frac{f(x+1)}{f(x)} = \frac{m_1^2 - (m_1 - m_2)\, x}{m_2\,(x+1)}$$

If this is inserted into the formula of $\varphi(x)$, we find:

$$\varphi(x) = \frac{x^2 - (2\,m_1 - 1)\, x + m_1^2 - m_1 - m_2}{x+1}$$

From this it will be seen that $\varphi(x)$ is different from 0. Some values of x will of course make $\varphi(x) = 0$, but this does not matter in this connection. We have, therefore, that the only value of c which will make $\alpha(x+c) = 1$, is $c = 0$. The approximation coefficient of the binomial frequency function is, therefore, a function of the position of the zero of the variable. In many cases the zero of the variable is known and in thoss cases the approximation coefficient found by Guldberg, can of course be used. But very often we will be faced with distributions where the zero of the variable is not known. This is for instance the case with distributions which are arranged from measurement data. In what follows I shall try to introduce an approximation coefficient of the exponential function which is independent of the value which is taken as the zero of the variable.

If we substitute $(x-k)$ and $(x+k)$ for x in the formula of the exponential frequency function (1), we find:

$$f(x-k) = f(x)\,.\,e^{\cdot\frac{2kx-k^2}{2m_2}}$$

and

$$f(x+k) = f(x)\cdot e^{-\frac{2kx+k^2}{2m_2}}$$

From this we find

$$f(x-k)\cdot f(x+k) = f^2(x)\cdot e^{-\frac{k^2}{m_2}}$$

and we can form the approximation coefficient

$$\alpha(x) = \frac{f(x-k)\cdot f(x+k)}{f^2(x)}\cdot e^{\frac{k^2}{m_2}} = 1 \qquad (2)$$

This approximation coefficient is independent of the position of the zero of the variable.

If a distribution is given in the form:

$$x = x_1 \qquad x_1 + l \qquad x_1 + 2\,l \qquad x_1 + 3\,l \quad \ldots\ldots \quad x_1 + n\,l$$

$$h(x) = h(x_1)\; h(x_1 + l)\; h(x_1 + 2\,l)\; h(x_1 + 3\,l) \ldots\ldots h(x_1 + n\,l)$$

where $(x_1 + i\,l)$ is the middle point in the class interval, we first calculate the values of the three middle moments of lowest order (m_1, m_2 and m_3). If $m_3 = 0$ (approximately), the distribution is nearly symmetrical. Further the values of the approximation coefficient are calculated by (2) wherein $f(x)$ has to be substituted by $h(x)$. If these values are approximately equal to unity, we conclude that the distribution in question can be replaced by the exponential frequency function.

The correctness of this conclusion is not directly obvious. We must also prove that no function other than the exponential frequency function satisfies the expression of the approximation coefficient (*).

This expression (2) can be written thus:

$$f^2(x) = f(x - k) \cdot f(x + k) \cdot e^{\frac{k^2}{m_2}} \qquad (3)$$

By taking the logarithms we find

$$2 \cdot \ln. f(x) = \ln. f(x - k) + \ln. f(x + k) + \frac{k^2}{m_2}$$

We now introduce the function

$$\varphi(x) = \ln. f(x)$$

and we get

$$2 \cdot \varphi(x) = \varphi(x - k) + \varphi(x + k) + \frac{k^2}{m_2}$$

x and k are here independent variables. If we now differentiate wit respect to k and later with respect to x, we get the two following equations:

$$0 = \frac{\partial \varphi(x + k)}{\partial k} + \frac{\partial \varphi(x - k)}{\partial k} + \frac{2\,k}{m_2}$$

$$2\,\frac{\partial \varphi(x)}{\partial x} = \frac{\partial \varphi(x + k)}{\partial x} + \frac{\partial \varphi(x - k)}{\partial x}$$

(*) The reader will observe that the following proof only deals with functions which can be differentiated two times at least.

We take the sum of these equations and, because

$$\frac{\partial \varphi (x - k)}{\partial k} = - \frac{\partial \varphi (x - k)}{\partial x},$$

we find:

$$2 \frac{\partial \varphi (x)}{\partial x} = 2 \frac{\partial \varphi (x + k)}{\partial k} + \frac{2 k}{m_2}$$

By differentiating with respect to k, ve get

$$0 = 2 \frac{\partial^2 \varphi (x + k)}{\partial k^2} + \frac{2}{m_2}$$

or

$$\frac{\partial^2 \varphi (x + k)}{\partial k^2} = - \frac{1}{m_2}$$

We nov introduce $z = x + k$ and because

$$\frac{\partial \varphi (x + k)}{\partial k} = \frac{d \varphi (z)}{d z}$$

ve have

$$\frac{d^2 \varphi (z)}{d z^2} = - \frac{1}{m_2}$$

If we now integrate this, we find

$$\varphi (z) = - \frac{(z - c m_2)^2}{2 m_2} + \frac{c^2 m_2 + 2 d}{2}$$

where c and d are integrating constants. Further we find:

$$f (z) = e^{\varphi (z)} = a \cdot e^{- \frac{(z - c m_2)^2}{2 m_2}}$$

where

$$a = e^{\frac{c^2 m_2 + 2 d}{2}}$$

We now introduce x again and find:

$$f (x) = a \cdot e^{- \frac{(x + k - c m_2)^2}{2 m_2}}$$

This is the mathematical formula of the exponential frequency function.

If we now form

$$f(x+k) = f(x) \cdot e^{-\frac{2k(x+k-cm_2)+k^2}{2m_2}}$$

and

$$f(x-k) = f(x) \cdot e^{-\frac{k^2-2k(x+k-cm_2)}{2m_2}}$$

we find

$$f(x-k) \cdot f(x+k) \cdot e^{\frac{k^2}{m_2}} = f^2(x)$$

The constants c and d are independent of k, buth they are not independent of each other.

It will be seen from this that the approximation coefficient $\alpha(x)$ given by (2) is a sufficient criterion to be used to prove whether or not a given distribution can be replaced by the exponential frequency function.

* * *

Guldberg (*l. c.*) uses the difference equation as a basis for the calculation of the theoretical frequency values. First he calculates the approximate probability of the most probable term. The probability of the other values of the variable are then calculated by means of the difference equation. The same method can also be used in the case of the exponential function. It is more simple, however, to calculate the theoretical frequencies in the following way.

By means of the equation

$$f(x-k) = f(x) \cdot e^{\frac{2kx-k^2}{2m_2}}$$

we can first calculate the proportion between $f(x-k)$ and $f(x)$ for all values of x. If we now take $f(x) = 1$ for a certain value of x, we can calculate relative values of $f(x)$. By taking the sum of these values and compare this sum with the actual number of observations, we find a factor which multiplied with these relative values, gives approximately correct values of $f(x)$ for all values of x. Theoretically the sum of the relative values should be taken over the whole range of exponential variation, *i . e .* from $x = -\infty$ to $x = +\infty$. In practice it is satisfactory to take the sum over the range of variation extending from $x = m_1 - 4\sqrt{m_2}$ to $x = m_1 + 4\sqrt{m_2}$.

We cannot find by this or by any other method the correct values of the theoretical frequencies because we must always use the calculated values (not the true values) of the moments.

In the following I shall consider some examples. Table 1 shows the distribution of Stature (in inches) for Adult Males born in England, Ireland, Scotland and Wales. (From the textbook of G. Udny Yule: An Introduction to the Theory of Statistics, p. 112. London 1912).

TABLE 1.

x	$h(x)$	$\alpha(x)$			$h'(x)$
		$k=1$	$k=2$	$k=3$	
57–58	2				0,8
58–59	4	2,04			2,4
59–60	14	0,97	1,55		8,7
60–61	41	0,80	0,74	1,83	27,1
61–62	83	1,17	1,47	1,51	72,7
62–63	169	1,33	1,76	1,89	167,4
63–64	394	0,85	0,97	1,26	331,5
64–65	669	1,01	0,85	0,96	677,3
65–66	990	0,97	0,98	0,83	991,6
66–67	1223	1,02	1,01	1,09	1248,0
67–68	1329	0,99	1,09	0,95	1350,3
68–69	1230	1,09	0,96	1,00	1256,1
69–70	1063	0,82	0,84	0,85	1004,8
70–71	646	1,16	1,09	0,98	691,5
71–72	392	0,99	1,00	1,00	408,4
72–73	202	0,88	0,93	1,62	207,5
73–74	79	1,20	1,84	2,02	90,7
74–75	32	1,44	1,80	2,98	34,1
75–76	16	0,73	1,13		11,0
76–77	5	1,49			3,1
77–78	2				0,7

The moments of this distribution are equal to (uncorrected):

$$m_1 = 67{,}5209\,,\quad m_2 = 6{,}6168 \text{ and } m_3 = -0{,}2088$$

As the third moment takes a very low value, the distribution is nearly symmetrical. The values of the approximation coefficient $\alpha(x)$ are calculated in the case of $k = 1$, $k = 2$ and $k = 3$ (see Table 1). It will be seen that these values are very nearly equal to unity, and we can conclude that the distribution can be replaced by the exponential frequency function. For the lowest and the highest values of x, the values of $\alpha(x)$ are rather high, especially in the case of $k = 3$. Some investigations which I should like to come back to at a later occasion, seem to show that the standard deviation of $\alpha(x)$ increases with increasing k. The most important values of $\alpha(x)$ are, therefore, those which are calculated with the lowest possible value of k, in this case with $k = 1$.

By means of the method described above, I have also calculated the theoretical frequency values. These are given in Table 1 under the head $h'(x)$. It will be seen from these also that there is a very close agreement between the distribution and the exponential function.

In « Matematisk Iagttagelseslære » (p. 126 Kobenhavn 1923) of J. F. Steffensen we find another distribution. This is here given in Table 2. x stands for the breadth (in millimetres) of the head of men, $h(x)$ for the number of men.

TABLE 2.

x	$h(x)$	$\alpha(x)$	$h'(x)$
120–124	1		
125–129	0		
130–134	0		
135–139	2		3
140–144	31	0,77	33
145–149	173	1,25	195
150–154	567	0,81	535
155–159	701	0,96	688
160–164	390	1,17	414
165–169	119	0,65	117
170–174	11	5,82	15
175–179	3	2,61	1
180–184	1	6,41	0
185–189	1		0
	2000		

The moments of this distribution are equal to (uncorrected) :

$$m_1 = 156{,}16\ ,\ m_2 = 32{,}9527 \text{ and } m_3 = +\ 20{,}4475$$

As the third moment is comparatively large, the distribution is a little skew. We have calculated the approximation coefficient in the case of $k = 5$ and the values are given i Table 2. It will be seen that these values deviate very little from unity for those values of x which are represented in high numbers. We must conclude from this that the distribution can be replaced by the exponential function with good approximation. Steffensen has calculated the theoretical frequencies and these are given in our table under the head $h'(x)$. These theoretical frequencies are calculated from the formula :

$$h'(x) = \frac{2000}{\sqrt{2\pi m_2}} e^{-\frac{(x-m_1)^2}{2m_2}}$$

wherein the uncorrected moments are used. It will be seen that the observed and the calculated frequencies agree very well.

In « Elemente der Exakten Erblichkeitslehre » (Jena 1913) of W. Johannsen we find a distribution which has been arranged by dr. C. G. Petersen. This distribution is given in Table 3. Here x stands for the number of finrays in the caudal fin of plaice and $h(x)$ stands for the frequencies.

TABLE 3.

x	$h(x)$	$\alpha(x)$	$h'(x)$
47	5		1,0
48	2	20,26	3,8
49	13	0,34	11,9
50	23	1,78	29,8
51	58	0,82	60,0
52	96	1,05	96,8
53	134	0,85	125,3
54	127	1,15	130,1
55	111	0,95	108,5
56	74	0,93	72,5
57	37	1,08	38,9
58	16	0,72	16,7
59	4	2,50	5,8
60	2	1,25	1,6
61	1		0,4
	703		

The moments of this distribution are:

$$m_1 = 53{,}67\ ,\ m_2 = 4{,}5392 \text{ and } m_3 = -0{,}7697$$

As the third moment has a very low value, the distribution is nearly symmetrical.

This distribution is discontinuous, and for that reason it may be argued that the distribution cannot be replaced by the continuous exponential function. In my opinion there is no great difference between a distribution like this and the continuous distributions. It is merely in the direct appearence we find the difference. This, however, is no place for a discussion of this question.

We have calculated the values of the approximation coefficient in the case of $k = 1$ (see Table 3). It will be seen that the values are approximately equal to unity. For $x = 48$ we find a very high value of the coefficient, but we see that this is only due to the relatively high value of h (47) and the low value of h (48).

By means of the method described above, I have calculated the theoretical frequency values. These values are given in Table 3 under the head $h'(x)$. It will be seen that there is a very close agreement between the observations and the exponential function.

In the Introduction I mentioned that skewness and irregularities in the distribution of measurement data are very often caused by lack in homogenity of the material. If we have a distribution of measurement data which does not agree with the exponential frequency function, it will very often be advantageous to arrange the material in groups according to some sign and see if the new distributions are exponential. In the following I shall deal with an actual example of this kind.

In Table 4 two distributions are given. The series headed $(A + B + C)$ shows the distribution of the length in *cm* of Norwegian large herring of the year-class 1904 caught in 1909. The moments of this distribution are (uncorrected)

$$m_1 = 29{,}07\ ,\ m_2 = 0{,}7770 \text{ and } m_3 = -0{,}2036$$

The values of the approximation coefficient $\alpha(x)$ are calculated (with $k = 1$), and given in Table 4.

It will be seen that the values of $\alpha(x)$ do not deviate very much from unity, but they have a remarkable decreasing tendency.

We must, therefore, conclude that the distribution does not agree well with the exponential frequency function.

The material at hand can now be arranged in three groups (A , B and C) according to signs which we shall not explain here (1).

TABLE 4.

$A + B + C$				A			
x	$h(x)$	$\alpha(x)$	$h'(x)$	x	$h(x)$	$\alpha(x)$	$h'(x)$
26	2		0,5	26	0		0,1
27	16	2,69	13,8	27	4		3,7
28	95	1,38	103,5	28	54	1,03	53,3
29	213	1,02	214,8	29	155	1,08	157,7
30	135	0,63	123,1	30	99	0,92	96,5
31	15		19,5	31	12		12,2
32	0		0,9	32	0		0,3
	476				**324**		

The distribution of length of the A group is also given in Table 4 (headed A). This distribution has the following moments

$$m_1 = 29{,}19 \text{ , } m_2 = 0{,}6343 \text{ and } m_3 = -0{,}0287$$

The third moment of this distribution has a much lower value than that of the $(A + B + C)$ — distribution. The first distribution is, therefore, more symmetrical than the latter. The values of $\alpha(x)$ for the A — distribution are very nearly equal to unity. The A — distribution, therefore, agrees very well with the exponential frequency function.

By means of the method described above, I have also calculated the theoretical frequency values. These are given i Table 4 under the head $h'(x)$. It will be seen that the observed and the theoretical frequencies agree better in the case of the A — distribution than in the case of the $(A + B + C)$ — distribution. But it will also be seen that even in the case of the latter distribution

(1) See about this : PER OTTESTAD. *Statistical Analysis of the Norwegian Herring Population*, « Rapp. et Proc. Verbaux », Copenhague, 1934.

there is no remarkable disagreement between the theoretical and the observed values. The approximation coefficient, however, shows very clearly that there are some disagreements between the distribution and the exponential function. The approximation coefficient $\alpha(x)$ is a very suitable indicator of the agreement or disagreement between observed distributions and the exponential frequency function.

The exponential frequency function of two variables.

The exponential frequency function of two variables has the same importance as a description of two — dimensional distributions as that of one variable in the case of one — dimensional distributions. In just the same way as we above have found an approximation coefficient of the exponential function of one variable, we can find approximation coefficients by means of which it is very simple to prove whether or not a distribution can be replaced by the exponential frequency function of two variables. First I shall remind the reader of some well known characteristics of this function.

Steffensen (*l. c.*) writes the mathematical expression of the exponential frequency function of two variables in the following way

$$f(x, y) = \frac{\sqrt{1-\rho^2}}{2\pi\sigma\tau}\, e^{-\frac{1}{2}\left(\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\tau} + \frac{y^2}{\tau^2}\right)} \qquad (5)$$

where the constants σ, τ and ρ are expressed by the moments in the formulae:

$$\left.\begin{aligned} \sigma &= \sqrt{\frac{m_{2x}\, m_{2y} - m_{1x}^2{}_{1y}^2}{m_{2y}}} \\ \tau &= \sqrt{\frac{m_{2x}\, m_{2y} - m_{1x}^2{}_{1y}^2}{m_{2x}}} \\ \rho &= \frac{m_{1x\,1y}}{\sqrt{m_{2x} \cdot m_{2y}}} \end{aligned}\right\} \qquad (6)$$

The latter constant (ρ) is called the coefficient of correlation. Very often it is used as a measure of the correlation between the two observed quantities upon which the correlation table (distribution)

is based. It should be remembered, however, that this constant can be calculated from every distribution of two variables by (6) were the theoretical moments are substituted by the moments calculated from the distribution in question. But it should not be used as a coefficient of correlation unless the distribution is exponential (1). It is therefore of the greatest importance to have a simple method by means of which it can be proved whether or not the distribution can be replaced by the exponential frequency function.

It will be seen that the partial integrals are equal to

$$F(y) = \int_{-\infty}^{+\infty} f(x, y)\, \partial x = \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}\,\tau} e^{-\frac{(1-\rho^2)\,y^2}{2\tau^2}}$$

and

$$F(x) = \int_{-\infty}^{+\infty} f(x, y)\, \partial \eta = \frac{\sqrt{1+\rho^2}}{\sqrt{2\pi}\,\sigma} e^{-\frac{(1-\rho^2)\,x^2}{2\sigma^2}}$$

or if σ, τ and ρ are expressed by the moments:

$$F(x) = \frac{1}{\sqrt{2\pi m_{2x}}} e^{-\frac{x^2}{2m_{2x}}} \text{ and } F(y) = \frac{1}{\sqrt{2\pi m_{2y}}} e^{-\frac{y^2}{2m_{2y}}}$$

If an observed distribution of x and y is exponential, the marginal distributions can be replaced by the exponential frequency function of one variable. When the problem is to find whether or not a two — dimensional distribution is exponential, it is most practical first to investigate the marginal distributions and see if these are exponential. If these distributions do not agree well with the exponential function of one variable, there is no reason to go further and see if the correlation table is exponential.

The first thing to do is, therefore, to calculate the values of the approximation coefficient $\alpha(x)$ and $\alpha(y)$ by (2) and see if these values are approximately equal to unity.

---

(1) The coefficient of correlation can also be used as a measure of the correlation if the distribution has such a form that it can be replaced by some other frequency functions. See about this: A. A. TSCHUPROW. *Grundbegriffe und Grundprobleme der Korrelationstheorie*. Berlin, 1925.

ALF GULDBERG: *On discontinuous frequency functions of two variables.* «Skandinavisk Aktuarietidskrift», Uppsala, 1934.

If we cut the functional surface (5) with the plane

$$y = \alpha x + \beta$$

we get an exponential frequency function of the form

$$f(x) = K \cdot e^{-\frac{1}{2}\left[\frac{1}{\sigma^2} - \frac{2\rho\alpha}{\sigma\tau} + \frac{\alpha^2}{\tau^2}\right]\left[x + \frac{(\sigma\alpha - \rho\tau)\sigma\beta}{\tau^2 - 2\rho\sigma\tau\alpha + \sigma^2\alpha^2}\right]^2}$$

where

$$K = \frac{\sqrt{1-\rho^2}}{2\pi\sigma\tau}\, e^{-\frac{1}{2}\frac{\beta^2(1-\rho^2)}{\tau^2 - 2\rho\sigma\tau\alpha + \sigma^2\alpha^2}}$$

The moments of this function are equal to

$$m_1 = -\frac{(\sigma\alpha - \rho\tau)\sigma\beta}{\tau^2 - 2\rho\sigma\tau\alpha + \sigma^2\alpha^2}$$

$$m_2 = \frac{\sigma^2\tau^2}{\tau^2 - 2\rho\sigma\tau\alpha + \sigma^2\alpha^2}$$

In the special case $\alpha = 0$, we find

$$m_1 = \frac{\rho\sigma}{\tau}\beta$$

$$m_2 = \sigma_2$$

Every frequency distribution of x according to every constant value of y has the same standard deviation equal to σ. It we substitute y for β in the expression for m_1, we get the equation of the first stright regression line:

$$m_{1x} = \frac{\rho\sigma}{\tau}\, y$$

We can also cut the functional surface (5) with the plane

$$x = \alpha_1 y + \beta_1$$

and we get the following function of y:

$$f(y) = K_1 \cdot e^{-\frac{1}{2}\left[\frac{\alpha_1^2}{\sigma^2} - \frac{2\rho\alpha_1}{\sigma\tau} + \frac{1}{\tau^2}\right]\left[y + \frac{(\tau\alpha_1 - \rho\sigma)\tau\beta_1}{\sigma^2 - 2\rho\sigma\tau\alpha_1 + \tau^2\alpha_1^2}\right]^2}$$

where

$$K_1 = \frac{\sqrt{1-\rho^2}}{2\pi\sigma\tau}\, e^{-\frac{1}{2}\frac{\beta_1^2(1-\rho^2)}{\sigma^2 - 2\rho\sigma\tau\alpha_1 + \tau^2\alpha_1^2}}$$

The moments are here:

$$m_1 = -\frac{(\tau\,\alpha_1 - \rho\,\sigma)\,\tau\,\beta_1}{\sigma^2 - 2\,\rho\,\sigma\,\tau\,\alpha_1 + |\,\tau^2\,\alpha_1{}^2}$$

$$m_1 = \frac{\sigma^2\,\tau^2}{\sigma^2 - 2\,\rho\,\sigma\,\tau\,\alpha_1 + |\,\tau^2\,\alpha_1{}^2}$$

In the special case $\alpha^1 = 0$, we find

$$m_1 = \frac{\rho\,\tau}{\sigma}\,\beta^1$$

$$m_2 = \tau^2$$

Every frequency distribution of y corresponding to every constant value of x has the same standard deviation equal to τ. If x is substituted for β_1 in the expression for m_1, we get the equation of the second regression line:

$$m_{1y} = \frac{\rho\,\tau}{\sigma}\,x$$

Every verical plane-cut in the functional surface is an exponential frequency function of one variable. This fact points out how to find a method for proving whether or not a given distribution can be replaced by the exponential frequency function of two variables. There can thus be found more approximation coefficients of the function. The following four coefficients are certainly of greatest importance.

$$P(x,y) = \frac{f(x-k,y)\,.\,f(x+k,y)}{f^2(x,y)}\,.\,e^{\frac{k^2}{\sigma^2}} = 1$$

$$Q(x,y) = \frac{f(x,y-l)\,.\,f(x,y+l)}{f^2(x,y)}\,.\,e^{\frac{l^2}{\tau^2}} = 1$$

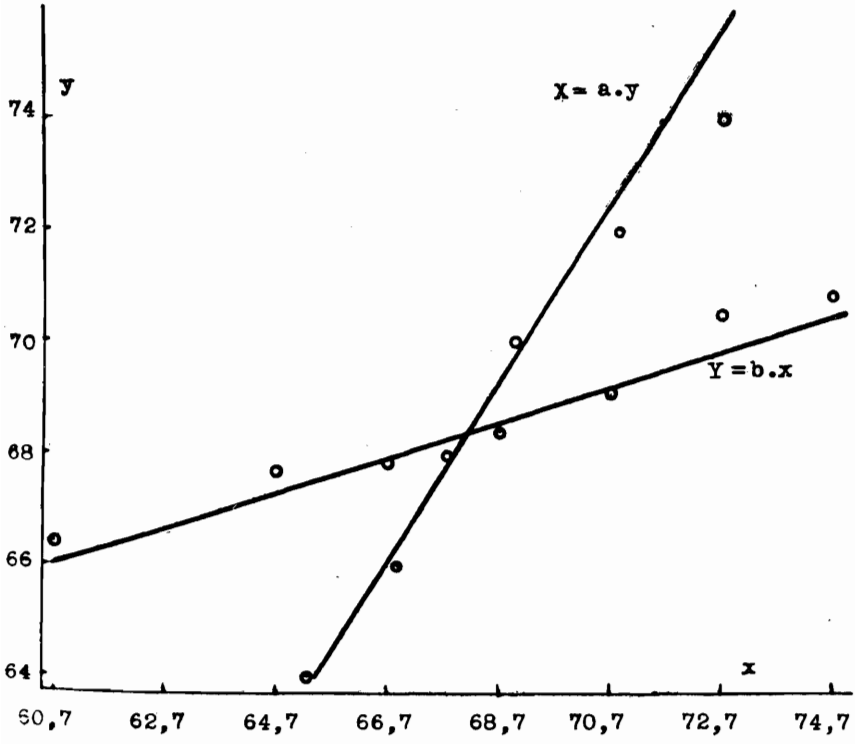
$$R(x,y) = \frac{f(x-k,y-l)\,.\,f(x+k,y+l)}{f^2(x,y)}\,.\,e^{\frac{k^2}{\sigma^2} - \frac{2\,\rho\,k\,.\,l}{\sigma\,\tau} + \frac{l^2}{\tau^2}} = 1$$

$$S(x,y) = \frac{f(x-k,y+l)\,.\,f(x+k,y-l)}{f^2(x,y)}\,.\,e^{\frac{k^2}{\sigma^2} + \frac{2\,\rho\,k\,.\,l}{\sigma\,\tau} + \frac{l^2}{\tau^2}} = 1$$

As mentioned we can also form some other coefficients. We can also take the proportion between two directly found coefficients. We have for instance:

$$\frac{P(x, y)}{Q(x, y)} = \frac{f(x-k, y) \cdot f(x+k, y)}{f(x, y-l) \cdot f(x, y+l)} \cdot e^{\frac{k^2}{\sigma^2} - \frac{l^2}{\tau^2}} = 1$$

It will be seen that if a distribution satisfies $P(x, y) = 1$ and $Q(x, y) = 1$ for all values of x and y, the proportion between



$P(x, y)$ and $Q(x, y)$ must be equal to unity. In cases, however, where there is some tendency in the disagreement between the distribution and the function, this tendency may be more remarkably expressed by the proportion between some of the approximation coefficients mentioned above. It is very difficult to present a definite rule for the actual work. The approximation coefficients which are most suited for the purpose in an actual case, will depend on the distribution itself. I will not go further into these questions, but will confine myself to one example to show how these approximation coefficients can be used.

The following example is taken from W. Johannsen (*l. c.* p. 357). In Table 5 stands for the height (in inches) of the full-grown children and y stands for the mean height of the parants (Galton's material).

TABLE 5.

y	x								
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7	
64	2	7	10	14	4				37
66	1	15	19	56	41	11	1		144
68	1	15	56	130	148	69	11		430
70	1	2	21	48	83	66	22	8	251
72			1	7	11	17	20	6	62
74							4		4
	5	**39**	**107**	**255**	**287**	**163**	**58**	**14**	**928**

The moments of this distribution have the following values:

$$m_{1x} = 68{,}0858 \qquad m_{1y} = 68{,}3642$$
$$m_{2x} = 6{,}6702 \qquad m_{2y} = 3{,}4321$$
$$m_{3x} = -0{,}2279 \qquad m_{2y} = -0{,}2208$$
$$m_{1x1y} = 2{,}1461$$

We will now first try whether the marginal distributions can be replaced by the exponential frequency function of one variable, and we calculate the values of the approximation coefficient $\alpha(x)$ and $\alpha(y)$. These values are given in Table 6. It will be seen that the third moment of both distribution has a very low value. Both distributions are, therefore, very nearly symmetrical.

TABLE 6.

x	$h(x)$	$\alpha(x)$	y	$h(y)$	$\alpha(y)$
60,7	5		64	37	
62,7	39	0,64	66	144	2,46
64,7	107	1,58	68	430	0,63
66,7	255	0,86	70	251	1,36
68,7	287	0,91	72	62	0,84
70,7	163	1,14	74	4	
72,7	58	1,24			
74,7	14				

For the x — distribution the values of the approximation coefficient are nearly equal to unity. This distribution can therefore be replaced by the exponential function. For the y — distribution also the values of the approximation coefficient are nearly equal to unity, but it will be seen that $\alpha\,(y)$ for $y = 66$, is rather high. This latter fact depends upon that $h\,(y) = 144$ is a little too low as we shall see later. The y — distribution does not deviate, however, very much from the exponential frequency function.

By inserting the calculated moments in formulae (6) we find the following values of the constants of the exponential function:

$$\sigma = 2{,}3083\,,\ \tau = 1{,}6558 \quad \text{and} \quad \rho = +\,0{,}4485$$

We have now calculated the values of $P\,(x\,,\,y)$, $Q\,(x\,,\,y)$, $R\,(x\,,\,y)$ and $S\,(x\,,\,y)$. The values of $P\,(x\,,\,y)$ are given in Table 7.

TABLE 7. $P\,(x\,,\,y)$.

y	x							
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7
64		0,86	2,08	0,43				
66		0,18	4,93	0,53	0,78	0,72		
68		0,53	1,32	1,04	0,87	0,72		
70		11,12	0,46	1,60	0,97	0,89	2,31	
72				0,48	2,08	1,61	0,54	

It will be seen that the value of $P\,(x\,,\,y)$ deviates considerably from unity for $x = 64{,}7$ and $y = 66$ and for $x = 62{,}7$ and $y = 70$. The latter disagreement does not matter so much. But the first shows that either $h\,(x\,,\,y)$ is too high for $x = 62{,}7$ and $y = 66$ and for $x = 66{,}7$ and $y = 66$ or $h\,(x\,,\,y)$ is too low for $x = 64{,}7$ and $y = 66$. The latter statement is the most probable, and we shall come back to this question later. The remaining values of $P\,(x\,,\,y)$ do not deviate very much from unity.

The values of $Q\,(x\,,\,y)$ are given in Table 8.

TABLE 8. $Q(x, y)$.

y	x							
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7
64								
66	8,60	2,01	6,67	2,50	1,51			
68	4,30	0,57	0,55	0,68	0,67	0,66	0,78	
70			0,55	1,70	1,02	1,16	1,95	
72							0,95	

It will be seen that all values of $Q(x, y)$ for $y = 66$ are greater than unity and that all values of $Q(x, y)$ for $y = 68$ are lower than unity except $Q(60{,}7, 68)$. This shows that either the values of $h(x, y)$ for $y = 64$ and $y = 68$ are too high or the values of $h(x, y)$ for $y = 66$ are too low. The same was found in the case of the marginal distribution of y. The remaining values of $Q(x, y)$ are approximately equal to unity. The high values of $Q(x, y)$ for $x = 60{,}7$ do not matter so much because the values of $h(x, y)$ are very low.

The values of $R(x, y)$ are given in Table 9.

TABLE 9. $R(x, y)$.

y	x							
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7
64								
66		1,76	8,90	1,67	2,03	1,28		
68		0,33	0,81	0,33	0,60	0,67	2,57	
70		0,88	0,84	0,94	1,13	2,40	3,02	
72						4,05		

We also see that the values of $R(x, y)$ are very high for $y = 66$. The value for $x = 64{,}7$ and $y = 66$ is remarkably high. The remaining values of $R(x, y)$ are satisfactory.

The values of $S(x, y)$ are given in Table 10.

TABLE 10. $(S\, x, y)$.

y	x							
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7
64								
66		1,03	13,55	1,66				
68		1,97	0,83	1,19	0,56	0,41		
70				1,50	1,63	0,65		

Here also we find a very high value for $x = 64{,}7$ and $y = 66$. The remaining values are approximately equal to unity.

From the values of the four approximation coefficients we can conclude that the exponential frequency function of two variables can be used as a good total discription of the distribution. There are some irregularities. But the only considerable deviation from the function that we find is for $x = 64{,}7$ and $y = 66$. The value of $h(x, y)$ for these values of x and y is too low. In this case it should be remembered that the total number of observations is not very high. There are, therefore, no reasons for believing that the deviations from the exponential function depend on factors other than those which are caused by the low number observations.

We can now also investigate the form of the regression curves. The straight regression lines of the exponential function are given by the formulae:

$$X = m_{1x} = \frac{\sigma \rho}{\tau}\, y = 0{,}63\, y$$

$$Y = m_{1y} = \frac{\tau \rho}{\sigma}\, x = 0{,}32\, x$$

These lines are drawn in Fig. 1. Here the means of y according to the various values of x and the means of x according to the various values of y are plotted. It will be seen that there is no remar-

kable disagreement between the theoretical and the empirical regression curves.

In the case of exponential distribution the standard deviations of the individual ditribution of x have the same value equal to σ, and the standard deviations of the individual distributions of y have the same value equal to τ. The standard deviations of the individual distributions of x and those of the individual distributions of y are given in Table 11.

TABLE 11.

x	σ_y	y	σ_x
60,7	1,60	64	2,12
62,7	1,68	66	2,24
64,7	1,75	68	2,21
66,7	1,75	70	2,48
68,7	1,55	72	2,27
70,7	1,54		
72,7	1,80		
$\tau = 1,66$		$\sigma = 2,31$	

It will be seen that the various values of σ_x are approximately the same and very nearly equal to σ, and the various values of σ_y are nearly the same and nearly equal to τ.

To show that the conclusion that this distribution is very nearly exponential, is correct I have calculated the theoretical frequencies by means of the method described above. The theoretical frequencies which are given in Table 12, are always given to the nearest whole number. If this table is compared with Table 5, it will be seen that there is a close agreement between the theoretical and the observed frequencies. From this it can also be seen that the greatest disagreement is found for $x = 64,7$ and $y = 66$. For these values of x and y the observed frequency is too low as we also found by means of the approximations coefficients.

TABLE 12. Theoretical frequency values.

y	x								
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7	
64	1	5	10	9	4	1			30
66	2	13	42	66	49	17	3		192
68	1	12	53	117	122	60	14	2	381
70		2	14	49	80	61	22	4	232
72			1	5	15	23	16	5	65
74					2	6	11	5	24
	4	**32**	**120**	**246**	**272**	**168**	**66**	**14**	**924**

Here I have used the calculated regression curves. It would perhaps be better to use the theoretical regression lines. It does not, however, make any great difference.

By DOUGLAS E. SCATES

Director of School Research, Cincinnati,

AND

Mrs. LUCILE MACK VAN NORTWICK

The influence of restrictive routes upon the center of minimum aggregate travel

The median of a two-way population distribution has received attention in the last few years (1) as the center around which the weighted deviations of the population locations total to a minimum (2). In this respect the median took over the property formerly (and erroneously) attributed to the mean (3), which is correctly the center around which the sum of the *squared* deviations is a minimum. *This two-dimensional* (4) *median however is not ordinarily given by the intersection of two orthogonal median lines*, and its determination has proved somewhat difficult. A practical approximation method has been suggested

(1) See the summarizing note by Frank Alexander Ross, *Editor's Note on the Center of Population and Point of Minimum Travel*, "Journal of the American Statistical Association", Vol. XXV (N. S.), p. 447-452, December, 1930.

(2) C. Gini e L. Galvani, *Di talune estensioni dei concetti di media ai caratteri qualitativi* (On Some Extensions of the Conceptions of Average to Qualitative Characters), "Metron", Vol. VIII, p. 3-209, July, 1929.

(3) Walter Crosby Eells, *A Mistaken Conception of the Center of Population*, " Journal of the American Statistical Association ", Vol. XXV (N. S.), p. 33-40, March., 1930.

(4) This terminology refers to the surface over which the population is distributed. In a strictly mathematical sense the varying population constitutes a third dimension.

by the writer and applied to the population of the United States (1). Another method has been used by Luigi Galvani to locate the median of the population of Italy (2).

Since this center of minimum aggregate travel is a point of practical interest, perhaps more than of theoretical interest, it seems desirable to derive techniques of calculating its location which will not depend upon assumptions that are clearly contrary to fact. The center, in its pure mathematical conception, assumes that travel will be in straight lines directly from the population locations to the center. In actual situations this condition can scarcely be approximated, except in ocean or air travel— or perhaps on unbroken plains. Such conditions do not obtain in the majority of the real problems, and the assumption is particularly untrue in cities.

It is the purpose of this paper to present methods of calculation which do not disregard actualities, and which are applicable to conditions where the routes of possible (or probable) travel are restricted in various ways. Two of these methods are illustrated and tested by application to a particular city--Cincinnati.

Four types of practical limitation upon travel will be discussed. The first, and perhaps the simplest, restricted situation is that where all travel must be along a line - - straight, curved, or broken— which connects the various population locations. This would be the case where people were to be brought together from a number of points along a seacoast, or alongside of a principal traffic route. The situation is not altered if the locations are off of the principal connecting route, and feed into it with branch routes, so long as the central meeting place is located on the

---

(1) DOUGLAS E. SCATES, *Locating the Median of the Population in the United States*, "Metron", Vol. XI, p. 49-65, June, 1933. Also published in "Proceedings of the International Congress for Studies on Population", Vol. X, Section of Methodology, p. 167-183. Rome, Istituto Poligrafico dello Stato, Libreria, 1934. Edited by Corrado Gini.

(2) LUIGI GALVANI, *Sulla determinazione del centro di gravità e del centro mediano di una popolazione, con applicazioni alla popolazione italiana censita il 1° dicembre* 1921, "Metron", Vol. XI, June, 1933. Also published in "Proceedings of the International Congress for Studies on Population", Vol. X, Section of Methodology, p. 121-152. Rome, Istituto Poligrafico dello Stato, Libreria, 1934. Edited by Corrado Gini.

route of travel. The principal condition is that there be but one route which can, or will, be used by the people to come together.

The treatment of this first case is the same as that of a linear frequency distribution. Even though the population locations may occupy a plane, and the travel arteries be ever so curved, the single connecting route makes them amenable to the treatment of a one-dimensional distribution and the median is calculated, along the route of travel, in the same manner as in an ordinary frequency distribution. The population locations become the observed values along the scale, and the populations become the frequencies. If the population points are off the main route, they are counted at the points where their branch routes enter the main route. Interpolation will of course not be attempted; the median either lies definitely at one of the population locations (or at the point where its branch route joins the main route), or it is indeterminate from one of these points to a second one (1).

A second type of situation occurs when the principal travel route is connected at the extremes- -does not have any end. This situation is somewhat less likely than the first, but it may occur where travel is around a body of water, or untraversed land, or where travel will be principally along trunk railroad lines connecting a number of major cities so as to form a closed figure.

Again, the situation can be reduced to that of finding the median as in a linear series, but in this case there would be as many different series as there are population locations, so that the solution is ordinarily not immediately obvious. There are relation-

---

(1) Griffin has published an interesting note dealing with cases in which the distribution of population is some mathematical function of the distance along which it is distributed. He includes seven cases of population located along a line where the median is to be on the line, in addition to an equal number of cases where the median is not specified as lying on the line, and also five cases of a population distributed over a plane. While the paper is primarily theoretical, the conditions assumed might in some cases be approximated. See F. L. GRIFFIN, *The Center of Population for Various Continuous Distributions of Population Over Areas of Various Shapes.* In " Proceedings of the International Congress for Studies on Population ", Vol. X, Section of Methodology, p. 153-157. Rome, Istituto Poligrafico dello Stato, Libreria, 1934. Edited by Corrado Gini.

ships that will aid one in picking out the most likely points by inspection, but there is no definite procedure available and the determination of the center of minimum travel will usually have to be based upon the summation of actual distances to several trial points. Cases of indeterminateness are likely (1).

From these two highly restricted, fairly simple situations we may go to the case of locating the median of an area in which travel is along routes laid out in rectangular fashion, such as streets frequently are in a city. This situation is particularly interesting, first, because it represents the condition obtaining in many practical problems, and, second, because it simplifies in a gratifying manner. It will readily be seen that, under the circumstances of rectangular travel, the distance from one point to another is the same regardless of the many optional routes which one may take, so long as one does not deliberately go out of his way. For example, the travel east and west between two obliquely located points will total the same whether one takes it all at once, or intersperses it with travel in the other direction in stair-step fashion, and the same may be said for the travel north and south.

This case may be handled simply and directly by two marginal frequency distributions, paralleling the two travel directions and representing the two linear aspects of the population distributed over the area. The median of each marginal distribution is then calculated, and the intersection of two median lines (projected from these median points, parallel to the lines of travel) is the center of minimum aggregate distance. Indeterminate areas or lines can result, but with real population weightings these are unlikely.

It should be noted that this condition can be generalized. The travel routes (distances) do not have to be along rectangular lines; it is only necessary that they lie parallel to two axes, which may be rhombic. Since the distance around any two consecutive sides of a parallelogram is equal to that around the other

---

(1) GINI has called " cyclical " this type of series (*Variabilità e Mutabilità*, Bologna, Cuppini, 1912), and he and Galvani have dealt thoroughly with the problem of the determination of the median for this type of series. For their treatment, see the reference cited previously in footnote 2.

two sides, and a whole is equal to the sum of its parts, the travel involved breaks down immediately into two linear aspects represented by the marginal distributions parallel to the lines of travel (1).

A fourth case arises when travel is likely to be along irregular arteries, with a number of interconnections, permitting many varied patterns of optional travel according as the point to be reached may shift slightly from one location to another. Even though oblique routes may be laid out systematically, as in the city of Washington, D. C., they must be considered irregular if they are not symmetrical around the real median. This condition again is a common practical one. It is referred to here as a maze. There is apparently no immediate mathematical attack for this problem, but it may ordinarily be handled without undue difficulty or tediousness by direct measurements. The procedure for solving this case will be illustrated in connection with the study of Cincinnati which follows. As in other cases, the median center may be indeterminate with respect to several nearby connecting points.

The actual conditions in a practical problem are frequently more complicated than in the single situations represented by any one of the foregoing cases. To test these procedures in an actual situation, and to compare the effect of various assumptions on the resulting location of the center, a study was made of the City of Cincinnati. This city was selected because of its marked irregularity in topography and in resultant travel routes. Further, data were available in terms of an excellent system of census tracts worked out by a number of civic bodies, especially the Department of Sociology at the University of Cincinnati, in conjunction with the United States Bureau of the Census. These 107 tracts afforded units of satisfactory size and number.

Three preliminary centers were calculated in the usual manner: the geographical (territorial) mean of the city, the mean of the population, and the intersection of orthogonal median

(1) Note that under the conditions given, the method of locating the median points by the intersection of two orthogonal median lines, discredited by recent literature, is actually valid. Under general conditions that method is valid only when the population is symmetrically distributed around the median in each direction, which means that the mean and median are the same in each of the directions used.

lines of population. For this purpose, the map was divided into small squares (about 200 of them for the city) and marginal frequency distributions made up through apportioning the area, or (1930) population of the census tracts, to these intervals as

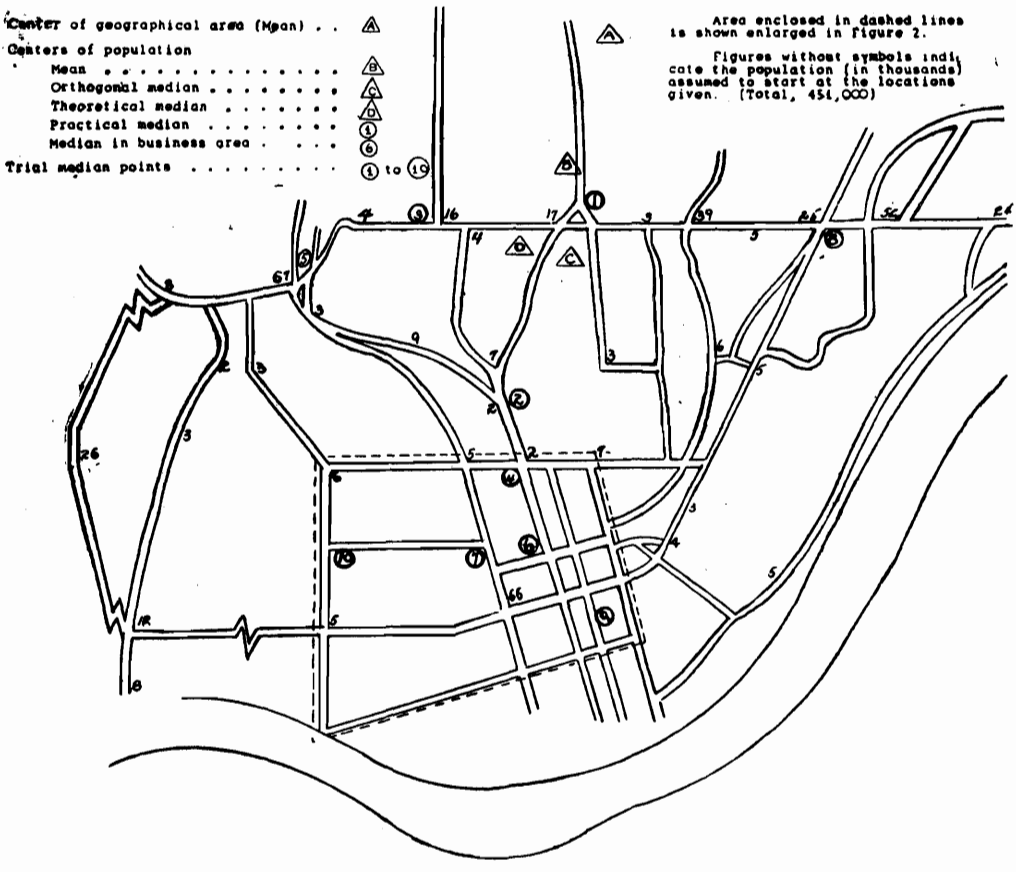


FIGURE 1. — Central Section of Cincinnati showing various centers and through travel routes.

required. Area required estimates only at the boundary of the city, and this was done in tenths of a square; population of separate census tracts was considered as though concentrated at the geographical center of each tract except in the case of a few territorially large tracts which were halved or quartered in advance. The resulting centers are shown in Figure 1. These centers afford a basis of comparison for the centers yet to be calculated.

For the purpose of locating the true median of the city's population, under theoretical conditions (i. e., the center from which the sum of the straight line distances to all of the popu-

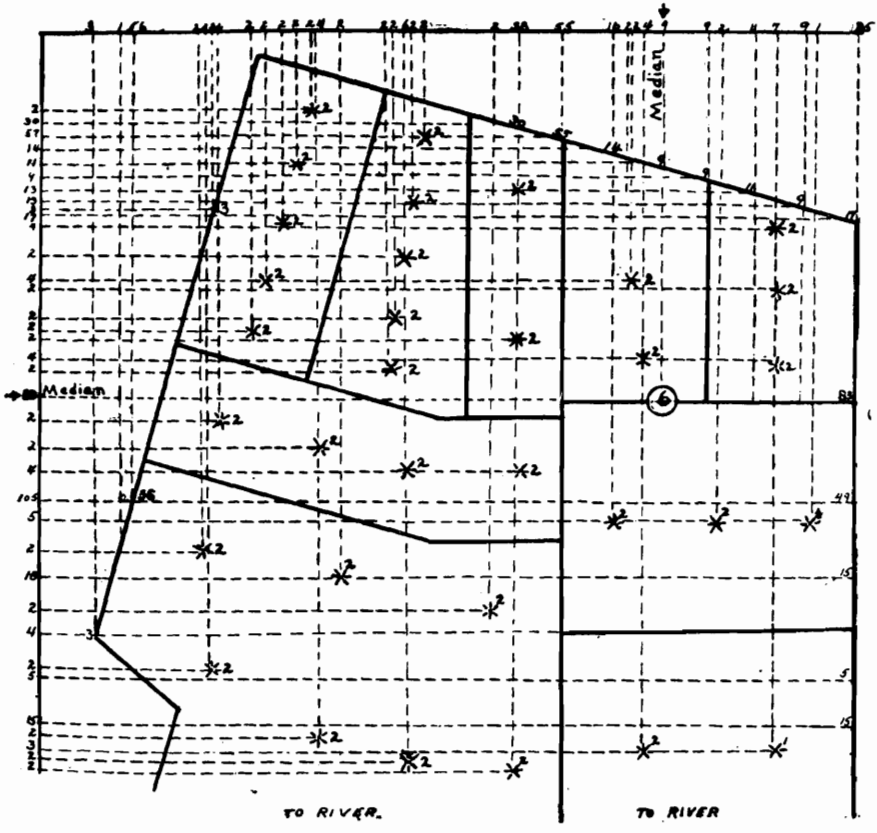


FIGURE 2. — Central business area permitting orthogonal travel, with marginal frequency distributions for calculating medians.

This section is shown in Figure 1 enclosed in dashed lines.

Figures in area show population (in thousands) at points where it is assumed to enter, or to originate within the area.

Axes are taken parallel to the orthogonal lines of travel in the section which contains the median, shown at (6)

Solid lines show boundaries of census tracts.

lation locations is a minimum), units having equal population are most convenient. Population units of four thousand each were decided upon as affording a satisfactory number and size of units. The census tracts were therefore combined, or divided

and apportioned as required to make up these new units of equal population, it being assumed in all cases where the contrary was not known that the population was evenly distributed over each census tract. The net error in this assumption is likely to be small with as many as 110 units ; nevertheless, because the errors are of increasing significance as they occur nearer the center, the fifty units near where the median was estimated to fall were all divided, yielding about 160 units for the city. Two, or if necessary, three different weights of units can be handled without undue difficulty if they lie systematically.

From these 160 units, or class intervals of population, the median for the city was calculated by a series of successive approximations. The procedure involved in this method is explained in detail in another place (1), and will not be repeated here. The resulting center is shown at D in Figure 1. While the assumption that travel will be in straight lines is, in Cincinnati, highly gratuitous, the proximity of this point to the practical median which is yet to be calculated, is interesting. It furnishes an example of how a distribution which is irregular in internal composition and travel routes may contain compensating elements which effectively neutralize each other.

The purely theoretical approach was abandoned at this point, and an attempt made to recognize the existing restrictions upon travel. An inspection of the layout of the city makes it clear that no single technique is likely to prove adequate. The main traffic arteries spread out fan-like from the business district and run out the valleys, so that an assumption of orthogonal or any two-directional travel routes would be generally inapplicable. The most heavily used routes are the most irregular ones. On the other hand, the area in and around the business centre is flat and is laid out in orthodox squares. The first attack on this problem was therefore to determine upon a central area (see Figures 1 and 2) within which travel was restricted only to the extent of being parallel to two axes, and apply to this area the technique described in connection with the third case.

The population lying outside of this central area can be brought within the scope of this technique by regarding the peo-

---

(1) SCATES, *op. cit.* The writer will supply reprints to those who indicate a desire for them.

ple from the outlying districts as situated along the perimeter of this central area at the points where they would normally enter it. That populations can thus be moved toward the center without affecting the determination of that center appears at first unreasonable, especially when the center is a function of distances. It will be recognized however that medians differ from moment measures in this very characteristic, and the outlying portions of the distribution, whether it be linear or areal, can be moved inward so long as their position with respect to the median is not altered. In the theoretical case, where travel routes are assumed to be direct, this procedure is not available since one does not know in advance the direction in which the median lies, and moving the population points would ordinarily alter their angular relation to it and affect its determination. Where restrictive travel routes are known, however, the population may be moved along the routes at will so long as the population is not moved past optional interconnecting routes which would affect the determination of the point of minimum aggregate travel. This latter contingency is dealt with later.

Determining the points of entrance into this area for the outlying population is primarily a matter of assigning the incoming traffic to the principal arteries. This can be done by several methods. It can be assumed that people will use the artery mearest to their residence, and the entire population outside the central area can be assigned on this basis. Or, figures on the street car and bus traffic on different streets can be obtained and used as a series of per cents for apportioning the entire population in the outlying districts. This is not satisfactory if there are heavy automobile arteries not paralleled by street car or bus lines; other assumptions also are involved. As a third basis, one may have traffic flow studies available--actual counts of vehicles using the different arteries.

In the present study the first and third bases were used. The entire outlying population was first assigned to the nearest arteries. The traffic flow maps (1) of the down-town area were then used to indicate how the traffic from any one of the large

(1) Maps were available from the office of the Regional Planning Commission of Hamilton County, Ohio, based upon recent traffic counts by the Cincinnati City Planning Commission.

arteries divides up as it approaches the business section, by crossing over from one route to another. Thus an estimate of the population entering the central area at each particular point was possible, and figures (in thousands) were accordingly set down along the perimeter of the area. The population dwelling within the central area was considered concentrated at the midpoints of small sub-areas. These internal midpoint figures and the perimeter figures were then projected to orthogonal axes to form marginal frequency distributions in two directions. The median of each distribution was calculated and projected back into the area and the intersection of these two median lines gave the median for the city. A detailed illustration of this procedure is shown in Figure 2.

This procedure is sound if several assumptions are sound. In the first place, the assumption is implicit that the real median of the city lies within the area that is chosen, for the procedure will produce a median whether or not the real median of the city is in the area. It should be noted however that if it be specified that a center must be found within the business area of the city, this technique is satisfactory regardless of the status of this assumption. Almost tantamount to the first assumption is the one that the central area is not so small as to cause the outlying traffic arbitrarily to be moved past interconnections that would permit a ready shifting of the traffic to reach a different central objective. Again, there is the assumption that travel within the area is restricted to two (or four, if you prefer) directions. Finally, there is the assumption that the outlying population can be accurately assigned to definite travel routes.

As it happens, none of these assumptions (with the possible exception of the last one, the status of which is unknown) is entirely met in the present instance. The danger of restricting the area too much was recognized, and the area was therefore made as large as possible, somewhat at the expense of the assumption concerning two directions of travel. The area was extended westward so as to include a large section which was properly laid out in squares, but which made a moderate angle with the streets of the main business section. This angle was almost entirely compensated for when the points were projected to the marginal distribution at the left, by making adjustments in the projection. It was felt better to include this additional area

than to assume initially that the median did not lie in it. The northern boundary of the central area was also extended somewhat on the basis of the same reasoning.

Still, it was suspected that the first assumption might not be true--the complex pattern of interconnecting arteries surrounding this area might place the median outside of the area if it were not artificially controlled when the area is limited to essentially orthogonal travel. The only remaining available technique was that mentioned in connection with the fourth case, referred to as the maze. This technique involves few assumptions; it requires only that travel be restricted to certain routes, that the population can be satisfactorily assigned to these routes, and that the research worker carry through his technique with reasonable judgment and persistence. In Cincinnati there is little doubt about the first condition, owing to the irregular topography, and the dominance of through streets. However if any side-streets would shorten the travel, they should be shown upon one's working map and used. The second condition is a purely technical one which probably involves no more error than the common grouping of data into class intervals. Some difficulty might however be encountered where automobile routes and street car or bus lines formed very different patterns.

This technique therefore appears to be satisfactory as a court of last resort, and as the only appropriate technique for a very irregular situation. It involves picking out a number of likely locations for the median and calculating the number of person-miles (or other population-distance units) required to reach each of these trial locations, finally choosing as the median the one which affords the minimum aggregate. It is essentially a cut-and-try method. The procedure however is not unduly arduous when properly systematized, and it offers the very great satisfaction of not requiring arbitrary assumptions. It is as flexible as the conditions may require, and, in carrying through the work, one exercises the same right of choice that the people would in picking their routes in actual travel, thus using the most direct routes.

As before, the outlying population is assigned to travel arteries, but it is not necessarily brought in so close to the center. It is recorded at the intersections farthest out which hold any likelihood at all of offering an advantageous change of routing

for any conceivable shift in the location of the median. The down-town population is regarded as concentrated at a few reasonable points--not so many being required as in the previous case. Altogether 33 population locations were used in the present study. These are shown in Figure 1, the population being indicated in thousands. The distance from each of these concentrated population locations to a selected trial point for the median is then stepped off with bow dividers set to represent some satisfactory unit, which may be entirely arbitrary. The number of such units of distance from the position of the population figure to the trial median is then multiplied by the appropriate population, giving the person-units from that location to the trial point. (It will be borne in mind that the distance which the people come beyond the place where the population figure is set down on the working map is entirely immaterial, since the number of person-units is important only for relative purposes). The sum of these products for the various population locations gives the aggregate person-units for that trial median. This process is repeated for each trial median until all of the reasonable choices have been exhausted, the trial point yielding the smallest aggregate being the median.

In stepping off the distances, a single count of the units of distance in a particular segment of route can serve all of the different population groups that will traverse that same segment of the travel route, so that the actual work can be greatly reduced. A little experience will enable one to select likely centers and to eliminate impossible ones very quickly. The choice is not haphazard, since definite indications quickly appear. In the present case ten trial medians were used, which is several more than would ordinarily be required; they were calculated as a matter of local interest to show the relative aggregate distance to a number of points in the business district. This supplementary information may, in many cases, be as valuable as is the location of the actual median itself. As a further matter of interest, the units converted were into miles; these are accurate for comparative purposes among themselves, but are not accurate in any sense as to absolute amounts. They cannot therefore be converted into per cents of the closest or farthest point tried.

The ten trial median points which were used are shown in circles in Figure 1, and the aggregate distances are given in

Table I. The results of this technique confirm the suspicion stated earlier that the median might not lie within the area of orthogonal travel as assumed by the preceding technique.

TABLE I. — *Comparative distance to ten different points. (Shown in circles in figure 1).*

TRIAL MEDIAN	Aggregate Person-Miles *	Average Miles * per Person
1. Vine and McMillan	777.85	1.723
2. Vine and McMicken	812.16	1.801
3. Clifton and McMillan	813.80	1.804
4. Vine and Liberty	821.09	1.821
5. Harrison and McMillan	832.60	1.846
6. Central Pkwy. and Race	847.17	1.878
7. Central Pkwy. and Laurel	900.28	1.996
8. Gilbert and McMillan	905.22	2.007
9. 5th and Walnut	1,007.91	2.235
10. Freeman and Laurel	1,104.50	2.449

(*) These are miles principally from the perimeter of the working area which was outlined so as assuredly to include the center of minimum aggregate travel. The average miles from residences would probably be two or three more.

The median just found is the center of minimum aggregate distance *when the actual existing travel routes are followed*, the most advantageous routes being used. Very interestingly, this median happens to be close to both the orthogonal and the theoretical medians (points *C* and *D* in figure 1). This juxtaposition is largely a matter of chance; if the cross street on which the practical median is located had been a mile or two farther out or closer into the city instead of where it is, this practical median would necessarily have moved considerably. The finding does, however, as an earlier comment stated, afford an interesting illustration of how a situation with considerable irregularity may contain compensating elements which cause it to

approximate perfect regularity. One would not, of course, ordinarily proced on the assumption of such a condition, for the odds are greatly against it.

In summary, the several techniques which have been discussed in this paper are all designed to locate the median, or point of minimum aggregate travel, under practical conditions. Four types of cases were discussed, the first two where travel is along a single route, the third where travel is parallel to two axes, and the fourth where travel is restricted to definite arteries which may present an entirely unsystematic pattern. The third technique was applied to the City of Cincinnati, illustrating the dangers which may lie in some of its assumptions. The fourth technique was also applied to Cincinnati, and while it may be the least satisfying from the standpoint of directness, in many situations it may be the only appropriate technique. These procedures locate, under the conditions given, the point at which all of the people in a certain territory could assemble with a mininimum of travel.

These techniques for locating practical medians recognize more restrictions than does a purely theoretical approach. It is true that some of them do not possess the degree of abstractness and generality that is so sought after by the pure mathematician, but it is also true that they do not necessitate assumptions which are not conceivably paralled by the facts. It may be observed in passing that the median never has been noted for a wealth of mathematical properties, and, while the median of an area may be defined in terms of central angles (1) when distances are taken directly, even this mathematical characteristic disappears when travel routes are restricted, leaving the single mathematical property by which the median is defined, namely, that the sum of the weighted deviations shall be a minimum.

This essential property of the median of an areal distribution (when properly determined), has only recently been recognized, and the median has not yet come into wide use. It does however offer many practical advantages. It is now clear that, in terms of practical significance, the median is a much more important center of areal distributions than is the mean. Amount

---

(1) This property was recently brought out by Gini and Galvani. See the references in footnotes 1 and 2.

of travel is at least one important factor in many decisions that are made. For example, in cities, techniques such as those here described would be serviceable in indicating the relative aggregate distance to various points that were being considered for the location of important civic or commercial centers. They should throw light on the need for improvement or creation of certain thoroughfares which may not at present be widely used, or existent. In cities where a particular school is to serve the entire city, or in rural regions where schools have to serve a considerable area, aggregate travel is important. In still larger areas manufacturers can utilize these techniques in deciding upon the location of large plants, using tons of raw material and of finished products for their weighting, in the place of population. In addition to the almost universal interest possessed by centers in general, the median has a peculiarly practical significance, and it should be widely utilized.

NORA FEDERICI

Ulteriori ricerche sugli studenti dell'Università di Roma

SOMMARIO: 1) Considerazioni generali e ricerche precedenti in argomento. — 2) I primi secoli di vita dello *Studium* e i laureati in *Utroque Jure* dal 1548 al 1729. — 3) Gli studenti iscritti all'Ateneo dal 1753 al 1815. — 4) La vita dello Studio dal 1815 al 1870. — 5) Lo sviluppo della Università dall'annessione di Roma al Regno d'Italia fino ai nostri giorni. — 6) La correlazione tra il numero degli studenti e la popolazione di Roma. — 7) La distribuzione degli studenti secondo le Facoltà. — 8) La distribuzione degli studenti italiani secondo la provenienza. — 9) La distribuzione degli studenti secondo la popolosità dei Comuni di provenienza. La cograduazione tra il numero degli studenti delle varie regioni e la distanza di queste da Roma. — 10) Gli studenti stranieri. — 11) La durata media degli studi per il complesso degli studenti e per gli studenti laziali.

È indubbia l'importanza e l'interesse di ricerche storiche che — documentate da materiale statistico — possano seguire lo sviluppo nei secoli delle varie Università italiane e che permettano, con l'ausilio appunto di tali dati, di illustrare le diverse fasi della loro vita. Soltanto uno studio statistico dà la possibilità di mettere in luce quale sia stata in realtà questa vita, in quale misura cioè le vicende storiche e gli ordinamenti degli studi abbiano influito sull'entità numerica della popolazione studentesca, sul suo orientamento scientifico e sulla sua composizione.

Già in questa stessa rivista sono stati pubblicati pregevoli studi in materia (1); in particolare uno studio sugli studenti dell'Ateneo romano attraverso il tempo è stato compiuto da Giorgio

(1) M. SAIBANTE, C. VIVARINI, G. VOGHERA, *Gli studenti dell'Università di Padova dalla fine del 500 ai nostri giorni*, in « Metron », vol. IV, n. 1. — C. PINGHINI, *La popolazione studentesca dell'Università di Ferrara dalle origini ai nostri tempi*, in « Metron », vol. VII, n. 1. — G. CAGNO, *Gli studenti*

Cagno (1), che si è proposto di esaminare i dati che egli ha potuto rintracciare sull'argomento, per seguire, sulla loro base, lo sviluppo dell'Università di Roma.

In seguito, in occasione dell'inaugurazione della Città degli Studi, altri dati sugli studenti del nostro Ateneo sono stati pubblicati da Nicola Spano, direttore amministrativo dell'Università (2).

L'Istituto di Statistica della R. Università, che aveva già promosso le ricerche del Cagno, ha fatto eseguire un'ulteriore ricerca nell'Archivio di Stato di Roma, con la speranza che potessero così essere colmate — almeno in parte — le lacune esistenti nel materiale finora pubblicato. Il risultato delle ricerche non è stato così completo come era da augurarsi, ma queste non sono riuscite inutili, in quanto è stato possibile mettere alla luce nuovo materiale, anche se non molto vasto.

Questo studio si propone di integrare, per quanto possibile, le ricerche sull'Università di Roma e, nello stesso tempo, di approfondirle mediante ulteriori elaborazioni, usufruendo a tal fine di tutti i dati disponibili, cioè anche di quelli pubblicati dal Cagno e, più recentemente, dallo Spano.

La rilevazione dei dati ancora inediti e parte delle elaborazioni sono dovute allo studente Luigi Farrace, che, come assegnatario di una borsa di Studio conferitagli dal predetto Istituto di Statistica, le ha compiute sotto la sorveglianza della scrivente, assistente presso la Facoltà di Scienze Statistiche, Demografiche ed Attuariali.

Per inquadrare il materiale nuovo, diremo anzitutto che quello pubblicato dal Cagno comprendeva : *a*) i laureati in *Utroque Jure* dal 1548 al 1729 ; *b*) gli iscritti alle varie facoltà dal 1813al 1870 ; *c*) gli iscritti (esclusi i fuori corso) e i laureati dal 1870 al 1930.

Dalla pubblicazione dello Spano risulta (oltre i dati già pubblicati dal Cagno : *a*) il numero complessivo degli iscritti dal 1789

---

dell' Università di Roma attraverso il tempo dal XVI secolo ai giorni nostri, in « Metron », vol. IX, n. 3-4. Cfr. anche G. LUCATELLO, *La popolazione studentesca dell' Università di Grenoble*, in « Metron », Vol. XI, n. 2. Questi studi intrapresi, come il presente, per suggerimento del Prof. Corrado Gini, come Direttore dell'Istituto di Statistica dell'Università di Padova, prima, e di quello della Università di Roma, poi, costituiscono l'inizio di attuazione di una più vasta serie di ricerche sulla popolazione studentesca dei vari paesi del mondo.

(1) G. CAGNO, cit.

(2) N. SPANO, *L' Università di Roma*. Casa Editrice « Mediterranea », Roma, 1935.

al 1812 e dal 1930 al 1934 ; *b*) il numero complessivo dei laureati dal 1831 al 1869 e dal 1930 al 1933.

I nostri dati comprendono : *a*) gli iscritti alle varie facoltà dal 1753 al 1773 ; *b*) le matricole dal 1788 al 1792 ; *c*) le lauree e gli altri gradi accademici per le varie facoltà : dal 1827 al 1840 per la facoltà giuridica e medica, dal 1832 al 1841 per la facoltà filosofico-matematica, dal 1833 al 1841 per la facoltà teologica.

Diciamo subito che i dati relativi alle « matricole » hanno indubbiamente lo stesso carattere di quelli degli iscritti ; questo deduciamo dall'esame di tali dati che mostra un ammontare di studenti simile a quello del periodo precedente da noi considerato e di quello successivo considerato dal Cagno. Non v'è quindi dubbio che la dizione di « matricole » avesse il significato generico di « iscritti » (1).

I dati relativi agli iscritti (compresi quelli delle matricole) possono utilmente venire inseriti in quelli del Cagno, poichè da essi risulta (come da quelli del Cagno) sia l'appartenenza alle varie facoltà, sia la provenienza degli studenti.

Per quanto riguarda, invece, i dati sui laureati, non crediamo opportuno inserirli in quelli già pubblicati e ci limitiamo a riportarli (tav. XI), senza farne oggetto di considerazioni di sorta, in quanto essi non si riferiscono all'anno in cui fu conseguita la laurea, bensì all'anno in cui fu spedito il diploma.

2. — Esula dai limiti imposti al nostro studio una sia pur sommaria trattazione delle vicende storiche dell'Ateneo romano (2). È però necessario ad esse continuamente richiamarci per illustrare lo sviluppo della popolazione studentesca dello *Studium Urbis* e per ricercare in esse le cause delle sue variazioni.

Per continuità di esposizione e per facilitare al lettore la visione

---

(1) A piena conferma della nostra interpretazione, sta la notizia riportata dallo SPANO, op. cit. pag. 52) di una disposizione del Rettore Mons. Luigi Costantini (1788), secondo la quale tutti gli studenti dovevano venire iscritti in una « matricola » che altro non era se non un registro dal quale risultavano tuti gli alunni iscritti all'Università e che poteva considerarsi come una specie di libretto di frequenza.

(2) Si sono ampiamente occupati della storia dello Studio romano il CARAFA (*De Gymnasio Romano*, Romae, 1751), in seguito il RENAZZI (*Storia dell' Università degli Studi di Roma*, Roma, 1803-1806) e, recentemente, lo SPANO (op. cit.).

comprensiva dell'andamento che presenta la popolazione studentesca del nostro Ateneo, comprenderemo nella nostra illustrazione tutto il materiale disponibile (v. grafico I). Ci atterremo, nel commento, all'esame obiettivo dei dati, ma tale esame limiteremo all'andamento generale, evitando di soffermarci sulle eventuali variazioni accidentali, che possono facilmente trarre in inganno.

Purtroppo, nulla ci è dato sapere sulla popolazione studentesca nei primi due secoli di vita dello Studio romano. Quando Bonifacio VIII lo fondò il 20 aprile 1303, la cultura era ancora, a Roma, retaggio quasi esclusivo degli ecclesiastici che studiavano le discipline sacre e le lettere nella « Schola Palatina » o « Universitas Romanae Curiae » che Onorio III, agli inizi del sec. XIII, sembra avesse fondato o almeno riorganizzato. Con l'apertura dello « Studio Generale » si inizia quindi una nuova êra per la diffusione della cultura, sopratutto fra i laici, anch'essi ammessi a far parte dello Studio. La mancanza di dati non ci permette però di stabilire con quale favore fosse accolta l'illuminata iniziativa del grande Pontefice. Le penose vicende della Corte Pontificia in quel periodo dovettero però indubbiamente favorire lo sviluppo dello Studio — sopratutto in vista del fatto che il trasferimento della Curia in Avignone privò Roma della « Schola Palatina » — se già nel 1318 Giovanni XXII concedeva ad esso la facoltà di conferire gradi accademici in diritto canonico e civile. Ma non sempre gli avvenimenti politici dovevano beneficamente influire sullo sviluppo dell'Università : gli scismi e le lotte politiche obbligarono i Pontefici a cure assai più gravi e l'Archiginnasio decadde fino al punto da venir chiuso più di una volta e per periodi anche lunghi. Ma l'Umanesimo e il Rinascimento non potevano mancare di farlo rifiorire. Infatti, lo Studio, riaperto nel 1406 da Innocenzo VII, dopo un altro breve periodo di decadenza, in seguito alle riforme di Eugenio IV (1) risorge a nuova vita; vi sono maestri, illustri umanisti, da Poggio Bracciolini a Lorenzo Valla, da Flavio Biondo a Pomponio Leto e a Francesco Filelfo.

---

(1) Sulla chiusura dell'Ateneo e sulla sua successiva riapertura da parte di Innocenzo VII e poi di Eugenio IV, sono d'accordo i più ; questa opinione della maggior parte degli storici è però confutata dal Valentini (R. VALENTINI, *Studium Urbis*, in « Archivio della R. Deputazione Romana di Storia Patria », vol. II, nuova serie, 1936), il quale giustifica le sue asserzioni con documenti degni di considerazione.

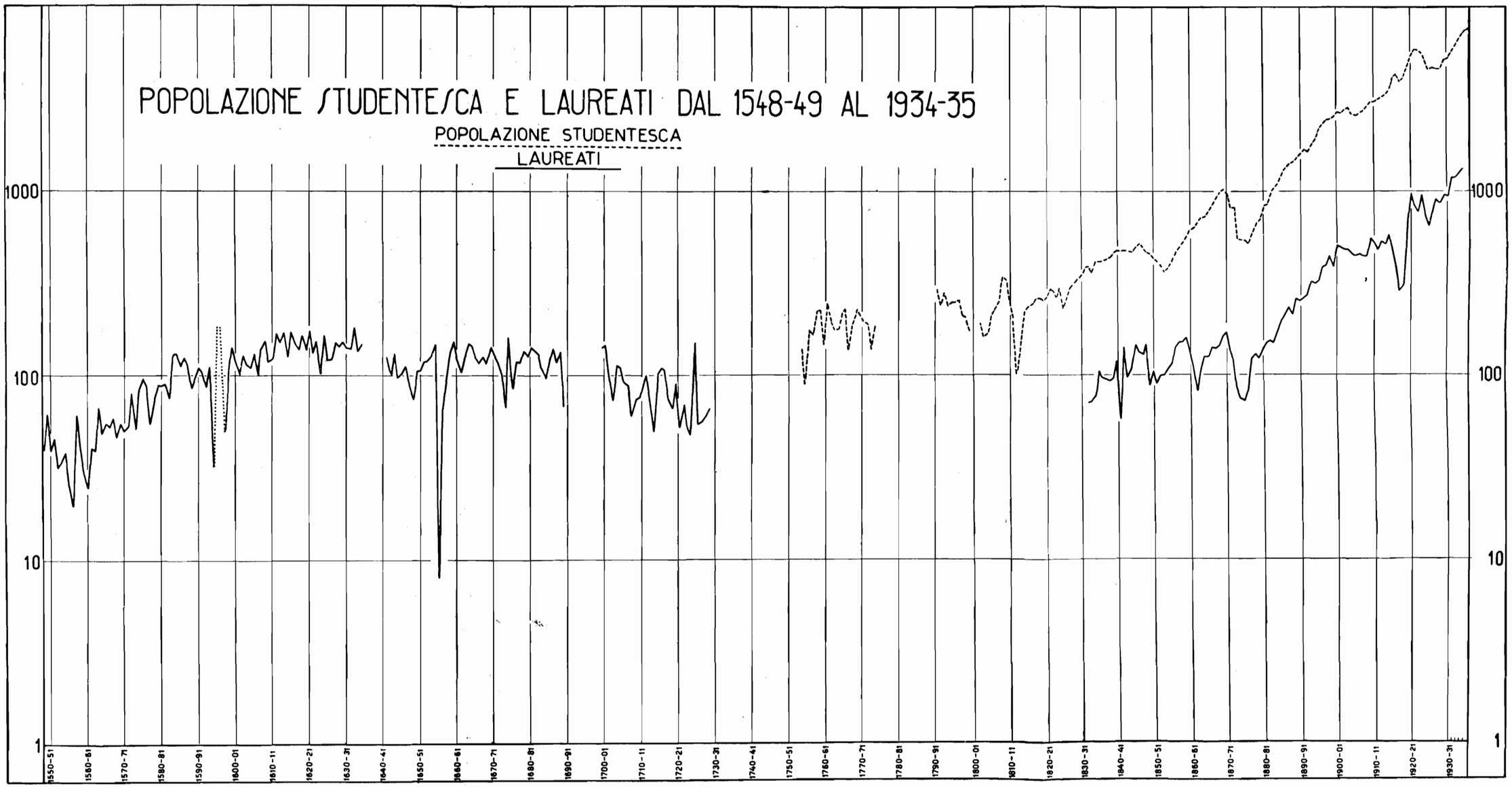


Grafico I.

Nota. — Nel presente grafico le ordinate sono proporzionali ai logaritmi dei numeri anzichè ai numeri degli studenti e dei laureati nei singoli anni.

Se lo Studio fiorì durante tutto il secolo, il periodo di massimo splendore fu sotto il pontificato di Leone X. Seppure nessuna notizia ci è dato conoscere circa l'entità numerica della popolazione studentesca di questo periodo, è certo però che essa aveva dovuto subire un notevole aumento se, in una costituzione pubblicata nel 1514, Leone X, con coscienza ed orgoglio profondamente romani, dice : « perchè l'Urbe così nelle lettere come nelle altre cose fosse capitale di tutto il mondo, è avvenuto che, nell'anno precedente, primo del nostro Pontificato, affluisse nella nostra città un numero così grande di studenti che già sembra il Ginnasio romano dover facilmente ottenere il primato su tutti gli altri d'Italia » (1). Le molteplici benemerenze del Pontefice umanista e il suo illuminato mecenatismo anche nei riguardi dello Studio romano furono tali che questo fu da alcuni chiamato l'Archiginnasio Leoniano.

Ma le previsioni di Leone X non dovevano avverarsi, chè la poca cura dei suoi successori e gli avvenimenti politici fecero decadere la « Sapienza » (è a quest'epoca che risale la denominazione ancor oggi familiare alla nostra Università) e, col Sacco di Roma, l'Ateneo fu addirittura chiuso.

Riaperto da Paolo III, che ne nominò « protettore » il nipote Alessandro Farnese, lo Studio rifiorì : particolare impulso fu dato alle scienze sperimentali, e delle provvidenze del Pontefice si avvantaggiò in special modo la Scuola medica, sopratutto per l'importanza data ai nascenti studi di anatomia.

A questo punto cominciano i primi dati : quelli dei laureati in *Utroque Jure,* pubblicati dal Cagno (2).

Il numero dei laureati si aggira, seppure con notevoli oscillazioni tra i 40 e i 50 ogni anno, fino al 1572-73. In seguito, si verifica un sensibile aumento, dovuto, presumibilmente, alle cure che all'Ateneo prodigò Gregorio XIII ; si deve infatti a questo Pontefice l'istituzione della cattedra di diritto criminale, il permesso dato ai neo-dottori di dare relazione dei loro studi personali, come pure il diritto loro riconosciuto di insegnare pubblicamente con la rimunerazione di 25 scudi all'anno, previa autorizzazione del Collegio dei Cardinali protettori e del Rettore.

L'incremento nel numero dei laureati si mantiene per tutto il secolo, salvo forti flessioni in qualche anno, come — ad esempio —

---

(1) Cfr. N. Spano, op. cit., pag. 18.

(2) Cfr. G. Cagno, cit., tav. I, pag. 168.

nel 1594-95 e nel 1597-98, in cui i laureati scendono rispettivamente a 32 e a 51. Queste cifre sono però con probabilità inesatte o per lo meno dubbie in quanto il Cagno le ha messe tra parentesi, senza però spiegare il significato della parentesi che noi — in mancanza di una spiegazione e in vista del sintomatico basso valore delle cifre — interpretiamo come indicazione di poca sicurezza dei dati.

I dati disponibili, riferendosi ai soli laureati in *Utroque Jure*, non permettono conclusioni definitive, ma pur tuttavia lo sviluppo dell'Università nell'ultimo trentennio del sec. XVI sembra indubbio (1).

Nè poteva tale sviluppo mancare se si pensi da un lato alle cure prodigate all'Ateneo dai Pontefici, dall'altro alla presenza di illustri maestri quali Paolo Giovio, Tommaso Inghirami, Pompeo Ugonio e Aldo Manuzio, e, nel campo delle scienze, Tommaso Giannozzi, Girolamo Fracastoro, Bartolomeo Eustachio, Andrea Cesalpino e Nicolò Copernico.

Nella prima metà del sec. XVII, il numero dei laureati si mantiene alto, salvo in qualche anno, in cui la diminuzione va probabilmente attribuita a condizioni particolari e contingenti, che non alterano però la regolarità generale dell'andamento. Una diminuzione sensibile si comincia a notare, invece, intorno al 1644-45 e persiste, seppure con una relativa ripresa negli anni 1645-46 e 1646-47, fino al 1650 ; il Cagno spiega la cifra di 76, relativa all'anno 1648-49, con la possibilità di una pestilenza importata dai profughi siciliani e napoletani in seguito all'insurrezione di Masaniello, ma il fatto che la diminuzione si fosse iniziata già quattro anni prima fa pensare che essa sia piuttosto dovuta — almeno in parte — a cause preesistenti, probabilmente in rapporto con le vicende politiche : la diminuzione è infatti conseguenza principalmente del minor numero di studenti provenienti dall'Italia Meridionale (2). Questa minore affluenza di studenti meridionali può forse essere attribuita allo stato di depressione economica e spirituale in cui si trovavano quelle popolazioni soggette alla Spagna, e che diede poi luogo alle insurrezioni di Napoli e di Messina.

---

(1) Diversa sembra l'opinione del Cagno, a quanto si può dedurre da qualche sua vaga osservazione in proposito ; ma l'osservazione appare ricavata più dall'esame singolo di qualche anno che da una visione complessiva del fenomeno. Cfr. G. CAGNO, cit., pag. 157.

(2) Cfr. G. CAGNO, cit., tav. II, pagg. 173-4.

Dopo un quinquennio di ripresa, il numero dei laureati scende a 8 nel 1655-56, in conseguenza di una gravissima pestilenza che, come lo stesso Segretario dello *Studium* avverte in una nota che il Cagno riporta, costrinse alla chiusura dell'Ateneo. Le conseguenze dell'epidemia dovettero farsi risentire a lungo, perchè, anche nei due anni successivi, i laureati sono relativamente pochi (rispettivamente 64 e 89) e solo nel 1658-59 il loro numero torna al livello normale e si mantiene tale per tutto il resto del secolo ; sensibili diminuzioni si notano sporadicamente in qualche anno (1673-74 ; 1675-76 ; 1683-84), diminuzioni che potrebbero essere del tutto accidentali, ma che potrebbero anche essere collegate a particolari condizioni politiche, giacchè esse generalmente risultano da una diminuzione nella percentuale sia — talvolta — di studenti esteri, sia — tal altra — di studenti nazionali extra-laziali e in particolare provenienti dall'Italia Settentrionale.

Particolare attenzione richiama invece la fortissima diminuzione del 1688-89, anno in cui il numero dei laureati scende a 68, da 135 che era nell'anno precedente ; dopo questo anno c'è una lacuna nei dati, che mancano per un decennio, fino, cioè, alla fine del secolo. Siamo all'epoca del Pontificato di Innocenzo XII che trovò l'Ateneo romano in fase di decadenza, nonostante che nel 1660 la nuova sede della « Sapienza » — solo allora ultimata dopo un secolo e mezzo di lavori — fosse stata solennemente inaugurata da Alessandro VII. Ma, nell'intervallo, le cure dei Pontefici per l'Ateneo si erano sempre più affievolite, tanto che questo decadeva lentamente ma progressivamente, ad eccezione forse della Scuola Medica, illustrata da docenti quali Marcello Malpighi e Giovanni Maria Lancisi. Verso la fine del secolo, la decadenza si accentua e, come si è visto, nel 1688 comincia a risentirne fortemente anche il numero dei laureati. Nè la mancanza di dati per gli anni successivi crediamo sia da attribuirsi con certezza allo smarrimento dei registri : grande doveva essere la disorganizzazione del servizio di segreteria e minimo il numero dei laureati, nè è da escludersi che questi siano perfino, in qualche anno, venuti a mancare, chè lo Studio era giunto a tale decadenza che Innocenzo XII fu sul punto di cederlo agli Scolopi, perchè ne facessero una scuola elementare gratuita per i figli del popolo. Fortunatamente questo malaugurato proposito non fu attuato, in grazia delle giustificate e vigorose proteste mosse contro di esso, che valsero a dissuadere il Pontefice. Questi, anzi, mutato disegno, si preoccupò di riformare

l'ordinamento universitario e di ovviare alle cause della decadenza dell'Archiginnasio. Le riforme furono progettate dal Cardinal Spinola, nominato Camerlengo, il quale le attuò energicamente. Tali riforme non dovrebbero aver mancato il loro effetto, almeno da quanto si può dedurre da una Memoria dell'epoca dove è detto che lo Studio « nel breve spazio di cinque anni erasi rimesso in prestigio, veggendosi le centinaia e centinaia di scolari non solo delle città circonvicine, ma ancora degli Studi più rinomati d'Italia e di fuori, frequentarlo giornalmente » (1). Se però ci riferiamo ai dati dei laureati in *Utroque Jure*, essi non sembrano confermare l'asserzione dell'anonimo autore della Memoria. Se, infatti, nei primissimi anni del secolo si nota una ripresa nel numero dei laureati, questo presto diminuisce nuovamente e, fino al 1728-29, anno in cui si arrestano i dati, solo eccezionalmente supera i 100.

Ciò è probabilmente da attribuirsi a cause di vario ordine. Innanzi tutto è probabile che « le centinaia e centinaia di scolari », ai quali allude l'Anonimo, fossero in gran parte studenti che abbiano frequentato lo Studio romano senza poi avervi conseguito la laurea. In secondo luogo non è improbabile che la Scuola legale — quella appunto alla quale si riferiscono i dati dei laureati — avesse meno risentito del beneficio della tanto lodata riforma dello Spinola. Questa seconda supposizione non parrebbe a prima vista accettabile, chè Clemente XI, al quale si deve di aver compiuto, sempre con l'ausilio dello Spinola, le riforme iniziate da Innocenzo XII, curò sopratutto la facoltà di giurisprudenza. Bisogna però tener presente un altro fatto : i Gesuiti avevano istituito una cattedra di diritto canonico, tanto frequentata da far sì che — appunto in quegli anni — sorgesse una controversia tra il Collegio Romano e l'Università, la quale ultima sosteneva di avere essa esclusivamente il diritto di tenere tale insegnamento, diritto che le fu effettivamente riconosciuto dal tribunale della Camera Apostolica e della Sacra Rota. La concorrenza che i Gesuiti facevano allo Studio può forse dunque aver bilanciato — per quanto si riferisce alla Scuola legale — gli effetti della riforma.

Poichè, però, gli studenti di giurisprudenza costituivano la percentuale maggiore, se questi, anzichè aumentare, mostrano piuttosto in questo periodo tendenza alla diminuzione (sia o meno

---

(1) *Memoria e memoriale a Clemente XI intorno allo stato antico e moderno dello Studio della Sapienza di Roma*, riportata anche dallo SPANO, op. cit., pag. 42.

questa da attribuirsi alla concorrenza dei Gesuiti) non rimane altro che concludere che l'aumentata frequenza di allievi dello Studio fosse in massima parte dovuta a studenti che non si siano poi presso di esso addottorati. Questo — s'intende — qualora si accettino senza riserve le asserzioni dell'Anonimo.

3. — A questo punto, come si è detto, cessano i dati del Cagno.

Per il ventennio 1753-1773 abbiamo i nostri dati (tav. I), dai quali risultano gli iscritti alle cinque Scuole: legale, medica, teologica, filosofico-matematica e di lingue.

Siamo nel periodo che va dalla fine del pontificato di Benedetto XIV alla fine di quello di Clemente XIV: l'Università dovrebbe dunque essere in pieno fiorire, dopo le nuove riforme del Cardinal Lambertini. Non è possibile fare confronti con periodi precedenti per i quali non abbiamo dati sugli iscritti, ma, se si consideri che in questo ventennio il loro numero si aggira intorno ad una media di 185, con un minimo di 90 nel 1754-55 ed un massimo di 244 nel 1760-61, dobbiamo concludere che la popolazione studentesca dell'Ateneo non doveva certo essere aumentata di molto. Infatti, nel primo trentennio del secolo, nel quale — abbiamo visto — il numero dei laureati era pure sensibilmente diminuito rispetto al secolo precedente, questo non è mai sceso al disotto dei 50 ed ha in qualche anno superato i 100, aggirandosi intorno ad una media di 80. Ora, pare certo alquanto strano che i laureati di una sola facoltà, seppure la più frequentata, possano raggiungere 1/3 del totale degli iscritti. Il numero medio di iscritti di questo ventennio è dunque presumibilmente inferiore a quello che lo Studio doveva contare nel primo trentennio del secolo. Si dovrebbe quindi concludere — con tutte le riserve che, in argomento di statistica storica, non sono mai troppe — che, nonostante le lodevoli e ripetute riforme, l'Ateneo non si fosse di molto avvantaggiato per quello che si riferisce all'entità numerica della massa studentesca.

Ulteriori dati sulla popolazione studentesca di questo periodo sono i nostri sulle « matricole », i quali vanno dal 1788-89 al 1792-93 (tav. II). Anche lo Spano riporta i dati della popolazione studentesca a partire dal 1789-90 (1); nei quattro anni comuni ai dati dello Spano e ai nostri si nota una qualche differenza nei valori, differenza però di poca entità. Continuando l'esame sui dati

(1) Cfr. N. SPANO, op. cit., pag. 313, segg.

da noi raccolti e, in seguito, su quelli dello Spano, notiamo come, in complesso, l'ultimo decennio del secolo ci mostri un numero di studenti più elevato di quello dei periodi precedentemente considerati : delle vicende politiche così poco favorevoli alla Corte Pontificia non ha dunque risentito l'andamento dell'Ateneo. Soltanto nel 1797 si comincia a notare un decremento nel numero degli studenti : questo continua nell'anno seguente e si accentua nel 1798-99 ; la diminuzione, che s'inizia, dunque, l'anno della proclamazione della Repubblica Cisalpina, diviene più sensibile quando anche Roma, occupata dai Francesi, viene proclamata Repubblica, e il Papa esiliato. La vita della Repubblica è breve ; i successivi rapidi sconvolgimenti politici obbligano alla chiusura dell'Ateneo, che viene poi riaperto da Pio VII il 26 novembre 1801 e riprende rapidamente vita. Le successive vicende di Roma e la sua annessione all'Impero Francese si fanno sì risentire sul numero degli iscritti all'Università, ma non in misura notevole, forse perchè l'incertezza politica era bilanciata dagli opportuni provvedimenti presi dal Governo repubblicano per uniformare l'ordinamento dell'Ateneo romano a quello delle altre Università dell'Impero.

Quando, nel 1814, tramonta l'astro napoleonico, la popolazione studentesca dell'Archiginnasio ammonta a 226 iscritti (1).

4. — Da questo momento si inizia un nuovo periodo storico per l'Italia e quindi anche la vita dello Studio romano muta il suo ritmo. D'ora innanzi, il numero degli studenti inizia un'ascesa che può considerarsi costante, se si eccettuino brevissimi periodi di arresto o di rallentamento in conseguenza di contingenti condizioni politiche ; tale ascesa, assai lenta in un primo tempo, interrotta negli anni immediatamente precedenti alla costituzione del Regno d'Italia, riprende poi, dopo il 1870, e continua fino ai nostri giorni con ritmo sempre più accelerato.

Ma torniamo a seguire i dati nel loro ordine cronologico.

Pio VII, oltre all'istituzione della Scuola degli Ingegneri, aumentò le cattedre portandole a 46 e, successivamente, provvide ad un riordinamento generale degli studi, proposto ed attuato dal Cardinal Consalvi. Seppure il nuovo ordinamento risentisse dell'inevitabile spirito reazionario, che in tutto lo Stato Pontificio si diffuse dopo il primo afflusso di idee liberali importate nel pe-

(1) Cfr. G. Cagno, tav. IV, pag. 186; N. Spano, op. cit., pag. 313.

riodo napoleonico, non può certo dirsi che il Consalvi mancasse di comprendere le nuove necessità didattiche determinate dall'evoluzione dei tempi, almeno per quel tanto che esse potevano armonizzarsi colla riaffermazione della supremazia della religione cattolica romana. E — in effetti — il numero degli iscritti mostra, nonostante qualche oscillazione, tendenza all'aumento dal 1814 al 1824, anno in cui raggiunge i 300.

Ma applicazione piena la riforma del Consalvi trovò soltanto sotto Leone XII, il quale, deciso a combattere con ogni severità il pericolo che le nuove idee poco ortodosse presentavano per lo Stato della Chiesa pose tutte le Università Pontificie (1), oltre che le altre scuole pubbliche e private, sotto l'assoluto e diretto controllo della Congregazione degli Studi. Forse a questo accentuato rigore reazionario è dovuta la notevole diminuzione che si nota nel numero degli studenti nel 1824-25 (235 contro 300 dell'anno precedente). Ma poco durò l'effetto sfavorevole della riforma Leoniana — la quale, del resto, aveva anche molto di buono — se già nel 1826-27 gli iscritti sono risaliti a 319 e il loro numero continua negli anni seguenti a crescere con sensibile regolarità. Nè turbano l'ascesa i provvedimenti severissimi di Gregorio XVI, che vedeva nell'Università uno dei più pericolosi centri di diffusione delle nuove idee che ormai cominciavano sempre più sensibilmente a farsi strada in Italia.

Con l'elezione di Pio IX, cessa il periodo reazionario e gli studenti entrano palesemente nella lotta politica : la diminuzione che si verifica nel loro numero in questo periodo è facilmente spiegabile con le gloriose lotte del nostro Risorgimento, alle quali i giovani parteciparono con un ardore che non ha bisogno di illustrazioni, e in particolare col fatto che le battaglie per l'Indipendenza, che si combattevano in tutta Italia, diminuirono l'affluenza a Roma dei giovani di altre regioni, più preoccupati di contribuire col loro sacrificio alla grande Causa che di affollare le aule universitarie. Intanto, Pio IX, pentitosi della sua primitiva politica, torna alla reazione e il numero degli studenti diminuisce ancora, fino a ri-

---

(1) Le Università Pontificie erano allora sette, due primarie : Roma e Bologna (la cui fondazione risale al sec. XI) e cinque secondarie : Ferrara (fondata nel 1391), Camerino (fondata nel 1727), Perugia (fondata nel 1308), Macerata (fondata nel 1540), Fermo e infine Urbino (fondata nel 1671 e ricostituita nel 1826).

dursi a 363 nel 1852-53, livello pari a quello di un ventennio prima. Ma breve è la sosta, chè lo sviluppo dell'Ateneo riprende subito e questo, raggiunti nuovamente i 500 iscritti nel 1857, giunge a contarne il doppio dopo soli dieci anni.

5. — Siamo ormai al 1870 (1) : i primi anni dopo l'annessione di Roma al Regno d'Italia sono anni di assestamento nè deve destar meraviglia se gli studenti diminuiscono. Chiusa la Facoltà di Teologia, per « sciopero » dei professori, venuta — almeno in gran parte — a mancare al nostro Ateneo tutta la massa degli ecclesiastici, che aveva costituito, sotto il dominio dei Papi, una percentuale certo non trascurabile della popolazione studentesca, inevitabilmente doveva — in un primo tempo — verificarsi una diminuzione.

Ma la diminuzione fu di breve durata e ben presto gli studenti crebbero in una proporzione sempre maggiore : 1043 già nel 1882-83 2210 dopo soli dodici anni, essi hanno toccato cifre sempre più alte e, con una evidente regolarità nel ritmo di accrescimento, solo turbata dagli sbalzi dovuti all'influenza esercitata dalla guerra mondiale, hanno raggiunto nel 1934-35 il ragguardevole numero di 7488, mentre le cifre provvisorie per il 1935-36 danno un ammontare di iscritti di 7743 (2).

Concludendo, possiamo affermare che il glorioso Ateneo romano, nonostante che periodi di oscurità si siano alternati a periodi di splendore, ha sempre resistito anche alle più avverse vicende politiche e si è rapidamente risollevato dopo i brevi momenti di decadenza ; esso, rinato a nuova vita con Roma capitale d'Italia, ha ormai conquistato un indiscusso primato sulle altre Università italiane.

---

(1) Cfr. G. CAGNO, cit., tav. VIII, pag. 196.

(2) Giacchè il Cagno aveva sempre escluso dal totale degli studenti quelli appartenenti alla Scuola di Ingegneria, all'Istituto Superiore di Scienze Economiche e Commerciali e all'Istituto Superiore di Magistero, anche la cifra relativa agli studenti del 1935-36 è stata da noi calcolata con gli stessi criterî, e questo per comparabilità del dato con i precedenti. Ma col nuovo ordinamento universitario (entrato in vigore nel 1935-36) anche questi tre istituti superiori sono entrati a far parte dell'Università, sì che, oggi, il numero complessivo dei suoi iscritti è di ben 11.437, cifra di gran lunga superiore a quella di tutte le altre Università d'Italia.

6. — Indubbiamente, lo sviluppo della popolazione studentesca dello *Studium Urbis* ha seguìto lo sviluppo della popolazione di Roma.

Vediamo come il numero degli studenti sia correlato con l'incremento demografico dell'Urbe.

Tralasciando il periodo per il quale abbiamo soltanto il numero dei laureati, abbiamo calcolato il coefficiente di correlazione tra la popolazione di Roma (1) e il numero degli iscritti all'Università, separatamente per il periodo 1756-1870 e per quello 1870-1930. Il coefficiente è risultato, per gli anni 1756-1870 = 0.633, valore dunque abbastanza rilevante; massima poi è la correlazione nel periodo 1870-1930, nel quale l'indice risulta = 0.988, qualora si tenga conto di tutti gli studenti, e = 0.995, qualora si considerino i soli studenti del Lazio.

7. — Terminato così l'esame sommario dell'andamento generale del numero degli studenti, veniamo a considerare qualche caratteristica della popolazione studentesca dell'Ateneo romano.

Sulla distribuzione degli studenti secondo le varie facoltà, poco si può osservare. I primi dati che ci permettano un tale esame sono quelli che si riferiscono al ventennio 1753-1773 (tav. III). Essi ci mostrano una costante e netta prevalenza della facoltà giuridica, i cui studenti non rappresentano mai meno del 70 % e superano spesso l'80 % del totale. La facoltà di medicina può considerarsi al secondo posto; per lo più però essa sta alla pari di quella di teologia e in qualche anno ne è anzi superata. Scarsamente popolata invece quella filosofico-matematica e ancora meno quella di lingue.

Già negli anni 1788-92 (tav. IV) la distribuzione per facoltà è mutata: figura sempre al primo posto la giurisprudenza, ma con percentuali assai più basse; sensibilmente maggiore risulta invece la frequenza nella facoltà di medicina — e sopratutto in quella di filosofia ed arti (2), che prevale sulla teologica (3) — ed aumentata

---

(1) I dati sulla popolazione di Roma sono stati tratti — per il periodo anteriore al 1870 — dalla *Monografia della città di Roma e della Campagna Romana*, pubblicata a cura della Direz. della Statistica Generale del Ministero dell'Agricoltura, nel 1881. Vedi, in particolare, vol. II, pagg. 353-56.

(2) Questa Scuola di « Filosofia ed Arti », così denominata nel 1788, corrisponde alla precedente Scuola « Filosofico-Matematica ».

(3) La Scuola teologica aveva, ora, assunto il nome di « Materie Sacre ».

pure nella facoltà di lingue. In complesso, in questi anni, la massa degli studenti risulta — in confronto al periodo precedentemente considerato — assai meno disugualmente distribuita tra le varie facoltà. Ciò è forse conseguenza del più razionale ordinamento delle Scuole, dovuto alle riforme di Mons. Costantini, confermate da Pio VI nel 1788.

Nel secolo seguente (1), fino al 1870, la distribuzione non appare mutata, salvo una più accentuata decadenza della facoltà di teologia, le cui percentuali scendono ora al disotto anche di quelle della facoltà di lingue ; essa è ormai l'ultima per frequenza di iscritti. Anche la nuova Scuola di Farmacia, seppure poco frequentata, conta, infatti, in quasi tutti gli anni, un numero di studenti superiore a quello della facoltà teologica.

Dopo la costituzione del Regno d'Italia, troviamo che la distribuzione degli studenti secondo le Facoltà non è — in complesso — variata (2) : infatti, le percentuali maggiori si riscontrano ancora per la Facoltà di Giurisprudenza, segue quella di Medicina e quindi quella di Scienze (Scienze Naturali, Matematica e Fisica, biennio propedeutico d'Ingegneria) la quale può *grosso modo* considerarsi la continuazione della Scuola filosofico-matematica. A distanza notevole seguono la Facoltà di Lettere (l'antica Scuola di Lingue, sebbene, qui, la differenza sia più radicale) e infine quella di Chimica-farmacia.

8. — Più interessante è l'esame della distribuzione degli studenti secondo la loro provenienza.

Vediamo quale appare questa distribuzione per i laureati in *Utroque Jure* di cui ai dati del Cagno (3).

Naturalmente, l'Italia Centrale è quella che dà maggior contributo alla popolazione studentesca dell'Ateneo, costituendone, salvo rarissime eccezioni, oltre il 50%. Tra gli studenti dell'Italia Centrale prevalgono i Laziali, ma tutt'altro che trascurabile è il numero degli extra-laziali. È ovvio che la maggior affluenza si

---

(1) Nel periodo 1793-1813, l'esame non è possibile, in quanto i dati (pubblicati dallo Spano) riguardano esclusivamente il totale degli iscritti e non permettono quindi la classificazione secondo la provenienza. Proseguiamo quindi l'esame sui dati degli iscritti riportati dal Cagno e relativi agli anni 1813-1870 : Cfr. G. CAGNO, cit., tav. VI, pag. 190.

(2) Cfr. G. CAGNO, cit., tav. XVII, pag. 212.

(3) Ibidem, tav. III, pag. 178.

riscontri da quelle regioni che appartenevano allo Stato Pontificio tanto che — tenuto conto di questo fatto — le percentuali dei laureati dell'Italia Centrale, in confronto con quelle, pure assai alte, relative all'Italia Meridionale non sembrano eccessivamente elevate. Questo trova forse una giustificazione nel fatto che lo Stato Pontificio contava nell'Italia Centrale — come già abbiamo visto (1) — altre quattro Università (Camerino, Perugia, Macerata e Fermo), esclusa quella di Urbino, che non ebbe vita continuativa. Queste Università, seppure di importanza assai minore in confronto allo Studio di Roma, dovevano necessariamente diminuire l'intensità dell'affluenza degli studenti all'Ateneo romano.

Notevolissima è la frequenza di studenti meridionali, spiegabile, sia con lo scarso numero e la deficiente organizzazione delle Università locali (2), sia anche con la maggiore vicinanza e facilità di comunicazioni del Sud con Roma.

Molto scarsa, invece, la percentuale di studenti settentrionali, nonostante che gran parte dell'Emilia fosse soggetta allo Stato Pontificio. Ma, certamente, le Università pontificie di Bologna e Ferrara (3) — e sopratutto la prima, primogenita e celeberrima tra le Università europee — dovevano assorbire la massima parte della popolazione studentesca di quella regione, mentre, d'altro lato, l'antichissimo e fiorente Studio di Padova, costituiva il centro d'attrazione di tutta la restante Italia Settentrionale.

Gran parte delle osservazioni finora fatte sembrano però perdere ogni valore quando si esamini la provenienza degli studenti iscritti al nostro Ateneo nel ventennio 1753-73 e nel quinquennio 1788-92 (tavv. V-VIII). Infatti, qui, le percentuali relative all'Italia Centrale sono sensibilmente più alte (mai inferiori al 70 %, raggiungono e superano in qualche anno l'80 %) ; notevolmente più

(1) Vedi nota 1 a pag. 103.

(2) A quell'epoca esistevano nell'Italia Meridionale soltanto tre Università : quella di Napoli (fondata come *Studium generale* da Federico II nel 1224) e, in Sicilia, quella di Catania (fondata da Alfonso d'Aragona nel 1434 e confermata con una bolla di Eugenio IV nel 1444) e quella di Messina (fondata da Paolo III nel 1548) ; queste Università e specialmente quella di Napoli — la più importante — lasciavano molto a desiderare nei riguardi della loro organizzazione. V. *L'Università di Napoli*, Ricciardi, Napoli, 1924.

(3) Bisogna, in ogni modo, tener presente che i nostri dati cominciano nel 1548, mentre parte dell'Emilia, tra cui Ferrara, passò sotto la completa giurisdizione dello Stato Pontificio solo alla fine del sec. XVI.

basse invece quelle degli studenti dell'Italia Meridionale. Soltanto per gli studenti settentrionali può dirsi che la loro proporzione sia rimasta, in media, invariata.

Potrebbe affacciarsi il dubbio che la più bassa proporzione che si nota per i laureati dell'Italia Centrale nei dati più sopra esaminati, in confronto di quella che presentano invece gli iscritti di questo periodo, sia da attribuirsi al fatto che gli studenti dell'Italia Centrale impiegassero un maggior numero di anni per compiere i loro studi. E l'osservazione sarebbe tutt'altro che inammissibile : non è illogico supporre che esista una maggiore selezione preventiva negli elementi forestieri che non in quelli locali, sopratutto quando si tenga conto del fatto che si tratta di epoca in cui le comunicazioni erano tutt'altro che agevoli e che quindi chi si accingeva a compiere gli studi fuori della propria città e spesso anzi fuori del proprio Stato doveva avere serie intenzioni di lavoro (1).

Non ci è possibile stabilire, per mancanza di materiale, se le variazioni riscontrate nella distribuzione regionale dei laureati del periodo 1548-1729 rispetto agli iscritti del periodo 1753-93 siano presumibilmente da attribuirsi al fattore sopra accennato o non piuttosto a variazioni effettive verificatesi nella proporzione degli studenti provenienti dalle diverse parti d'Italia. L'unica possibile pietra di paragone è data dai dati sui laureati che abbiamo del periodo 1827-41, periodo non troppo lontano da quello in esame. Questi dati però, come si è detto, non sono attendibili, in quanto si riferiscono agli anni di spedizione del diploma. Ma nostro scopo è, qui, soltanto di stabilire se la variazione nella distribuzione regionale degli studenti si verifichi anche per i laureati e per tale scopo il materiale è utilizzabile e ci mostra come — effettivamente — anche nei laureati si noti un'altissima percentuale per l'Italia Centrale, mentre assai bassa è quella per l'Italia Meridionale. Sembra perciò da escludersi che le variazioni riscontrate nei dati degli iscritti del periodo 1753-93 siano da attribuirsi, almeno in misura preponderante, ad una più lunga durata degli studi per i giovani dell'Italia Centrale.

La diversa proporzione nelle percentuali deve piuttosto

---

(1) La possibilità di una minor selezione preventiva degli studenti laziali in confronto di quelli extra-laziali è confermata dai risultati del calcolo della durata media degli studi, che appare effettivamente più lunga per gli studenti del Lazio. V. pag. 117 e grafico II.

ritenersi, almeno in buona parte, conseguenza del diminuito numero di studenti meridionali. Quale la causa? Non è facile individuarla con certezza, ma si può ammettere che le cure prodigate allo Studio napoletano da Carlo III di Borbone, il Principe riformatore, e sopratutto, forse, la presenza in quel periodo di un maestro quale Antonio Genovesi, abbiano richiamato i giovani delle provincie limitrofe, riducendone l'esodo verso Roma (1).

Nei nostri dati del periodo 1753-93 figurano anche gli studenti della Corsica (2), che abbiamo creduto opportuno mantenere distinti per poter esaminare sia il loro ammontare, sia le variazioni in esso verificatesi in questo periodo, il quale presenta particolare interesse per un tale esame. Infatti la cessione della Corsica alla Francia, avvenuta nel 1768, potrebbe aver determinato una diminuzione nel numero degli studenti côrsi che compivano gli studi a Roma. Le percentuali ci mostrano come assai rilevante fosse l'affluenza all'Ateneo romano dei giovani côrsi (in media quasi pari a quella dei provenienti dall'Italia Settentrionale o dall'Italia Meridionale e di gran lunga superiore a quella dei provenienti dall'Italia Insulare) e — quel che è assai importante — come tale affluenza non fosse affatto diminuita dopo la cessione dell'Isola alla Francia.

Continuando il nostro esame sui dati degli iscritti dal 1813 al 1869, riportati dal Cagno, vediamo che le proporzioni nelle percentuali della provenienza si mantengono analoghe a quelle del periodo precedente (3), fino — almeno — al 1857, anno in cui comincia a verificarsi una progressiva diminuzione delle percentuali relative all'Italia Centrale, accompagnata da un corrispondente aumento tanto di quelle dell'Italia Settentrionale che dell'Italia Meridionale. È questo un evidente frutto delle condizioni politiche: la coscienza italiana comincia a farsi sentire in tutti; Roma è, nel

---

(1) Al Genovesi, eccezionale maestro dell'Università di Napoli, che per primo osò «dettare» in italiano, furono attribuiti 15.000 discepoli. Vedi: *L'Università di Napoli*, cit., pag. 456.

(2) Nei dati del Cagno non compaiono mai studenti della Corsica; sembrandoci inammissibile che soltanto negli anni ai quali si riferiscono i nostri dati frequentassero l'Ateneo romano studenti côrsi (tanto più che in questi anni essi risultano assai numerosi) pensiamo che il Cagno non abbia distinto i Côrsi dagli altri Italiani. Non è però facile individuare in quale classe li abbia compresi, tanto più che nella sua classificazione non figura nemmeno l'Italia Insulare, ma soltanto l'Italia Settentrionale, Centrale e Meridionale.

(3) G. Cagno, cit., tav. VII, pag. 192.

cuore di tutti, la Capitale d'Italia, anche se in realtà dovrà passare ancora più di un decennio perchè il sogno si avveri, e gli studenti di tutta Italia cominciano ad affluire più numerosi a Roma, meta di tutte le speranze e di tutti i desideri.

Col 1870, il fenomeno si accentua (1) : Roma Capitale attrae sempre più i giovani, e gli studenti dell'Italia Centrale non rappresentano più che il 60-70 % dell'intera massa studentesca. Quanto più ci avviciniamo agli anni a noi più prossimi, tanto più vediamo affluire a Roma numerosi i giovani : oggi oltre il 40 % degli iscritti all'Ateneo romano è costituito da studenti dell'Italia Settentrionale e Meridionale, più numerosi questi ultimi (oltre il 30 %), meno i primi (circa il 10 %).

La distribuzione degli studenti secondo la provenienza non varia di molto se, invece di considerare gli iscritti, consideriamo i laureati di questo periodo per i quali pure abbiamo i dati.

9. — Sulla distribuzione degli studenti secondo la loro provenienza sarebbe stato interessante fare altre indagini, ma il materiale di cui disponevamo non ce lo ha permesso, chè i dati già pubblicati non sono suscettibili di classificazioni diverse da quelle fatte dagli autori. Il materiale nostro, d'altra parte, è troppo scarso e frammentario.

Abbiamo, in ogni modo, voluto, a scopo di saggio, esaminare la distribuzione degli studenti a seconda della popolosità del Comune di provenienza e la cograduazione fra il numero degli studenti e la distanza da Roma delle varie regioni. Entrambi i calcoli sono stati fatti soltanto per gli studenti iscritti nel periodo 1753-1773.

Nei riguardi della prima ricerca (tavv. IX-X), abbiamo suddiviso i Comuni di provenienza in quattro classi : Comuni con oltre 50.000 ab. ; Comuni con 20 — 50.000 ab. ; Comuni con 10 — 20.000 ab. ; Comuni con meno di 10.000 ab. (2).

Le percentuali più alte in media risultano quelle relative alla prima classe, il che è ovvio in quanto questa comprende gli studenti di Roma. Seguono immediatamente le percentuali relative ai Comuni con meno di 10.000 ab. e, a distanza, quelle relative

(1) G. Cagno, cit., tav. XXI, pag. 224.

(2) La popolazione dei Comuni considerata è quella risultante al censimento del 1871.

alle altre due classi di Comuni. Qualora si escludano gli studenti di Roma dalla prima classe, questa presenta percentuali in media analoghe a quelle della seconda e terza classe di Comuni. Si può pertanto concludere che i Comuni minori (quelli con popolazione inferiore ai 10.000 ab.) danno maggior contributo di studenti allo *Studium Urbis*, seguiti, a distanza presso a poco uguale, dalle altre tre classi, ove si eccettui la perturbazione che sulla prima classe esercita l'inclusione di Roma.

Abbiamo voluto considerare la possibile influenza determinata dalla distanza sul maggiore o minor numero degli studenti provenienti dai vari compartimenti. A tal fine abbiamo calcolato, come si è detto, l'indice di cograduazione. I vari compartimenti sono stati graduati tenendo conto della distanza da Roma delle rispettive città principali. Evidentemente, il criterio è abbastanza imperfetto, tanto più che le distanze considerate sono quelle attuali, mentre le vie di comunicazione del tempo erano certo diverse. Ma non era facile stabilire un criterio di graduazione migliore; basterà perciò tener conto che l'indice ottenuto risponde soltanto *grosso modo* al valore della cograduazione.

L'indice è risultato = —0.536 qualora si considerino tutte le regioni e = — 0.225 qualora invece si escluda il Lazio. Il primo valore è fortemente influenzato dalla prevalenza di studenti romani; il secondo indica una contrograduazione assai scarsa, nè questo deve destar meraviglia, chè certamente altre numerose circostanze dovevano contribuire alla maggiore o minore affluenza degli studenti all'Ateneo romano e principalmente l'appartenere o meno allo Stato Pontificio, e il diverso grado di organizzazione degli studi nelle varie regioni, poichè dove fiorivano Studi di fama, assai minore era certamente l'emigrazione di giovani studiosi.

Una ricerca sull'influenza esercitata dalla distanza sul numero degli studenti che dalle varie regioni si recano nelle diverse Università è stata fatta dal Castrilli (1). Tale ricerca è diversa dalla nostra, oltre che per l'epoca (si riferisce al triennio 1911/12-1913/14), anche perchè in essa è stata considerata — per ogni regione — la cograduazione tra le distanze che separano il capoluogo di regione da ognuna delle varie Università e il numero degli studenti che tali Università frequentano. Il calcolo è stato eseguito

(1) V. CASTRILLI, *Gli studenti delle Università italiane*, Roma, Tip. Operaia Romana, 1923.

distinguendo le varie facoltà. Se noi facciamo le medie degli indici ottenuti dal Castrilli (1) per le varie regioni, otteniamo i seguenti risultati (2):

	Metodo *a*	Metodo *b*
Lettere	— 0.69	— 0.61
Giurisprudenza.	— 0.53	— 0.46
Scienze e Ingegneria	— 0.40	— 0.28
Medicina	— 0.61	— 0.57
Veterinaria	— 0.83	— 0.36
Farmacia	— 0.91	— 0.61

Tali risultati mostrano come la distribuzione degli studenti nelle varie Università sia notevolmente cograduata con le distanze che li separano da dette Università.

Ma questi risultati, da una parte si riferiscono all'epoca contemporanea, in cui le comunicazioni sono uniformemente più facili e la soppressione delle frontiere politiche tra regione e regione italiana ha eliminato uno dei maggiori fattori di perturbazione; dall'altra non escludono che tale cograduazione possa, ancor oggi, risultare minore per qualche Università, che eserciti sugli studenti una particolare attrazione, la quale venga a diminuire, se non ad annullare, l'influenza della distanza.

Questo potrebbe essere il caso dell'Università di Roma e così potrebbe spiegarsi il valore notevolmente più basso (— 0.225) da noi ottenuto calcolando — come si è detto — la cograduazione tra il numero degli studenti delle varie regioni e le distanze che separano da Roma i capoluoghi delle regioni stesse.

Del resto, un'altra ricerca del Castrilli confermerebbe questa ipotesi e quindi il risultato da noi ottenuto. Il Castrilli ha infatti trovato che alcune Università più importanti — e tra queste Roma — esercitano sugli studenti un'attrazione più forte di quella che sarebbe da attendersi dal criterio della distanza (3).

---

(1) Cfr. V. CASTRILLI, cit. pag. 150.

(2) Il Castrilli ha calcolato gli indici con due metodi diversi: nel metodo *a* egli ha considerato tutte le Università esistenti; nel metodo *b* egli ha invece compreso solo le Università verso le quali si dirigevano effettivamente gli studenti di ciascuna provincia.

Nella pubblicazione del Castrilli, gli indici compaiono col segno + in quanto il calcolo è stato eseguito tra il numero degli studenti e l'inverso delle distanze.

(3) Cfr. V. CASTRILLI, cit., pagg. 154-156.

10. — Un esame a parte merita la frequenza nell'Ateneo romano degli studenti stranieri.

Seguiremo il loro andamento in base alle percentuali annue calcolate sul totale degli iscritti. Dato il numero relativamente scarso di studenti stranieri, specialmente nei secoli passati, non è facile individuare la tendenza dell'andamento, in quanto le percentuali subiscono variazioni fortissime di natura evidentemente accidentale. Ci limiteremo perciò ad un esame a grandi linee per la difficoltà di distinguere le variazioni accidentali da quelle causate da ragioni estrinseche.

Per il periodo più antico abbiamo, come ripetutamente si è visto, solo i dati dei laureati (1).

Fino al 1600 la proporzione di stranieri, pur subendo forti oscillazioni da anno ad anno, si mantiene tendenzialmente costante, aggirandosi intorno ad una media del 4 %. Nel sec. XVII, invece, si comincia a verificare una tendenza all'aumento, che, nonostante molte oscillazioni, si accentua fino al 1650, anno in cui abbiamo un massimo relativo nella percentuale degli studenti stranieri i quali raggiungono il 12.2%. Successivamente, per circa un ventennio, si nota una diminuzione alla quale segue una forte ripresa, che culmina nel 1686-87 con un valore della percentuale del 16.8 %, (massimo di tutto il periodo considerato) e che, nonostante una lacuna nei dati, possiamo supporre si mantenga per tutto il secolo, finchè, nel 1701-02, comincia di nuovo l'andamento decrescente e le percentuali, all'infuori di oscillazioni accidentali, si mantengono, nel primo trentennio del sec. XVIII, intorno ad una media del 5-6 %.

Le nazionalità che più frequentemente e con maggior continuità ricorrono nel periodo che abbiamo finora considerato sono la francese, la polacca, la spagnola e la tedesca. In particolare, noteremo che l'incremento degli studenti stranieri, verificatosi nella prima metà del sec. XVII è sopratutto dovuto al maggior numero di francesi accorsi al nostro Ateneo, che in seguito, invece, diventano più scarsi. Il massimo delle percentuali che abbiamo notato nell'ultimo trentennio del secolo è dovuto, invece, in gran parte ad una temporanea, notevolissima affluenza di Olandesi (limitata al ventennio 1670-90) e ad una accentuazione della frequenza di Belgi e di Tedeschi.

---

(1) G. Cagno, cit., tavv. II, III, pag. 170, 178.

Esaminato questo primo periodo, possiamo, per i secoli successivi, seguire l'andamento degli studenti stranieri sui dati degli iscritti.

Il ventennio 1753-73 (tavv. V-VI) è caratterizzato da una notevolissima affluenza di studenti stranieri. Le percentuali di questo periodo sono assai alte e si aggirano intorno ad una media dell'8 %.

Prevalgono numericamente di gran lunga, sugli altri, i Tedeschi; seguono a distanza gli Spagnoli e i Polacchi, mentre le altre nazionalità sono rappresentate solo scarsamente e saltuariamente.

I tre periodi di maggiore frequenza di stranieri nella nostra Università possono forse spiegarsi con le condizioni di risveglio politico e spirituale, e quindi culturale, dei Paesi che a tale frequenza maggiormente contribuiscono. La prima metà del secolo XVII è un periodo di floridezza per la Francia sotto il governo del Richelieu ; la fine di quel secolo corrisponde invece al massimo splendore per l'Olanda che, affermatasi come potenza marittima, fiorisce sotto Guglielmo d'Orange nel campo del pensiero e della arte ; la seconda metà del secolo XVIII è invece il periodo di maggior potenza della Prussia che si afferma nel campo politico sotto Federico il Grande.

Continuando il nostro esame, vediamo che, nel quinquennio 1788-92, (tavv. VII-VIII) le percentuali si mantengono ancora alte ma presentano già una diminuzione rispetto al ventennio precedentemente considerato.

In seguito, i dati del Cagno, che vanno dal 1813 al 1869 (1), mostrano una notevole diminuzione degli stranieri, diminuzione che si accentua dopo il 1830 e, ancor più, dopo il 1850. Le nazionalità più rappresentate in questo periodo sono, nell'ordine, la francese, la spagnola e la svizzera (2).

Dopo il 1870 (3), la costituzione del Regno d'Italia ha arrestato per un quinquennio l'affluenza degli stranieri, che presto però riprendono a frequentare la nostra Università in numero sempre maggiore, sebbene il rapido e crescente incremento della popolazione

---

(1) G. CAGNO, cit., tavv. V, VII, pagg. 188, 192.

(2) Non possiamo fare a meno di notare come — fatto assai strano — nei dati del Cagno di questo periodo non compaiano affatto studenti tedeschi.

(3) G. CAGNO, cit., tavv. XIX, XXI, pagg. 218, 224.

studentesca nazionale mantenga le loro percentuali ad un livello assai basso (2 % circa). Oggi, numerosi stranieri di tutte le nazionalità europee ed extra-europee vengono a studiare nel nostro Ateneo : i mezzi di comunicazione odierni e gli scambi culturali internazionali permettono ai giovani dei più lontani paesi di affluire a Roma.

11. — Un'ultima ricerca abbiamo fatto per indagare come sia variata nel tempo la durata media degli studi.

Per tale esame ci siamo valsi del numero degli iscritti e di quello dei laureati (1).

Per ottenere la durata media degli studi, bisognerebbe dividere il numero degli iscritti per il numero dei laureati che da essi provengono. Qualora la durata dei corsi fosse uguale per tutte le facoltà e non vi fossero nè studenti fuori-corso, nè studenti che conseguono abbreviazioni di corso, i laureati — in ogni anno — proverrebbero tutti dagli iscritti di uno stesso anno. Il rapporto andrebbe allora istituito tra gli iscritti di un dato anno e i laureati di tanti anni dopo quanti sarebbero quelli necessari per conseguire la laurea. Un'influenza perturbatrice, però, è determinata dalla variazione del numero degli iscritti : per ridurre al minimo tale influenza, abbiamo raggruppato i dati in periodi decennali. Per la scelta dell'ampiezza dello scarto tra gli anni corrispondenti in cui considerare iscritti e laureati, ci siamo basati sulla regolarità dei rapporti che si ottenevano tenendo conto di intervalli di diversa ampiezza, e su tale base abbiamo scelto l'intervallo di quattro anni (2).

Se esaminiamo l'andamento dei rapporti (v. grafico II) per il complesso degli studenti, vediamo che essi si aggirano dapprima intorno ai 4 anni, discendono quindi al disotto di questo valore toccando un minimo di 3.4 nel decennio 1850-59 e, dopo tre anni, riprendono a crescere, fino a raggiungere un massimo di

---

(1) La ricerca comprende il periodo 1835-1935, perchè soltanto per tale periodo disponevamo tanto del numero degli iscritti quanto di quello dei laureati.

(2) I rapporti sono stati calcolati nel modo seguente : il primo rapporto si è ottenuto dividendo la somma del numero degli iscritti dal 1835-36 al 1844-45 per la somma dei laureati dal 1839-40 al 1848-49 ; il secondo : dividendo la somma del numero degli iscritti dal 1836-37 al 1845-46 per quella dei laureati dal 1840-41 al 1849-50, e così via.

7 nel decennio 1866-75, per poi discendere nuovamente e quindi risalire, mantenendosi, per tutto questo periodo, intorno ad una media di 5-6. Una tendenza alla diminuzione si comincia di nuovo a verificare negli ultimi anni.

Si deve quindi concludere che il profitto degli studi — nei riguardi almeno della loro durata — è, in complesso, buono: l'influenza di elementi provenienti da altre Università, che si iscrivono

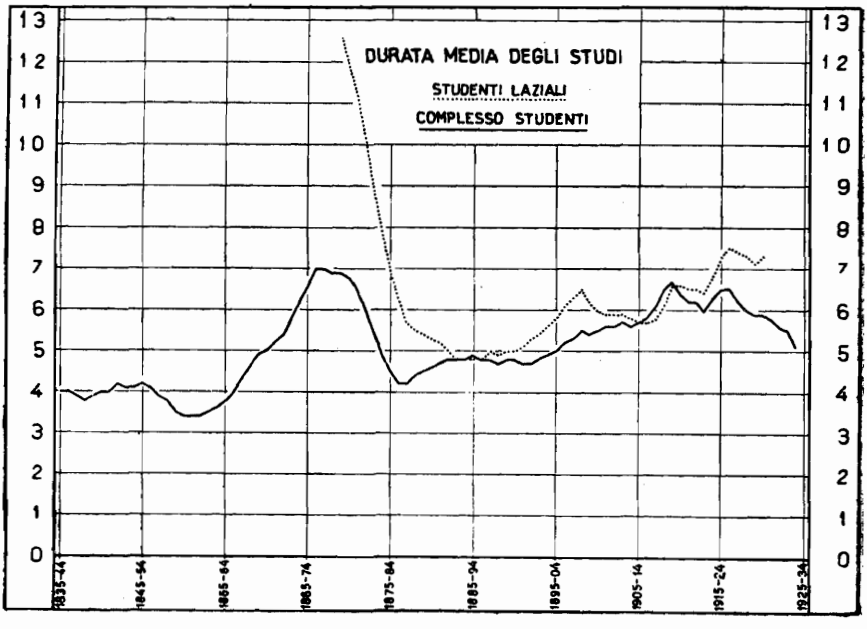


Grafico II

ad anni di corso diversi dal primo, è particolarmente sensibile in qualche periodo, tanto da determinare la discesa dei rapporti al disotto di 4.

L'alto valore dei rapporti che si riscontra dal 1860 al 1874 corrisponde al periodo immediatamente precedente ed immediatamente seguente alla costituzione del Regno d'Italia, periodo di particolare perturbazione per l'Italia e in special modo per Roma, nel quale, con molta probabilità, numerosi studenti debbono aver

abbandonato gli studi prima di aver conseguito la laurea. Le altre minori irregolarità che si presentano sono — crediamo — più che altro di carattere accidentale.

Maggiore appare in media la durata degli studi per gli studenti del Lazio (1).

Eccezionalmente alti sono i rapporti, che a questi si riferiscono, dal 1870 al 1875, e ciò probabilmente in relazione alle conseguenze dell'annessione di Roma al Regno d'Italia: molti studenti iscritti all'Ateneo (e in particolare gli ecclesiastici) avranno con probabilità interrotto gli studi, molti saranno passati ad altre Università. Intorno al 1880 i rapporti cominciano ad assumere valori normali e si aggirano, con oscillazioni più o meno accentuate, intorno ad una media di 6, mantenendosi quasi costantemente al disopra di quelli ottenuti per il complesso degli studenti. In corrispondenza degli anni di guerra si nota un aumento, che manca invece quasi completamente per il complesso degli studenti.

Il fatto che la durata media degli studi sia leggermente più elevata per gli studenti del Lazio in confronto del totale degli studenti, si può spiegare con l'osservazione, già fatta più sopra, di una minor selezione preventiva dei giovani che studiano sul luogo, rispetto a quelli che si spostano da località più o meno lontane.

* * *

L'incremento numerico degli studenti negli ultimi anni ci fa sperare che l'antico e glorioso Studio romano, oggi rinnovato nella sua sede con la costruzione della Città degli Studi, seguirà nella sua progressiva ascensione lo stesso ritmo che il Fascismo ha felicemente impresso allo sviluppo dell'Urbe.

---

(1) Il periodo considerato va, per questi, dal 1870 al 1930.

BIBLIOGRAFIA

Annuario degli Istituti Scientifici Italiani diretto dal Prof. SILVIO PIVANO, Roma, Athenaeum, 1919.

G. CAGNO, *Gli studenti dell'Università di Roma attraverso il tempo dal XVI secolo ai giorni nostri*, in « Metron », vol. IX, n. 3-4.

G. CARAFA, *De Gymnasio Romano*, Romae, 1751.

V. CASTRILLI, *Gli studenti delle Università italiane*, Roma, tip. Operaia Romana, 1923.

G. LUCATELLO, *La popolazione studentesca dell'Università di Grenoble*, in « Metron », vol. XI, n. 2.

L'Università di Napoli. Napoli, Ricciardi, 1924.

Monografia della città di Roma e della campagna romana. Direz. della Statistica Generale, Ministero Agricoltura, Industria e Commercio, Roma, 1881.

C. PINGHINI, *La popolazione studentesca dell'Università di Ferrara dalle origini ai nostri tempi*, in « Metron », vol. VII, n. 1.

F. M. RENAZZI, *Storia dell'Università degli Studi di Roma*, Roma 1803-1806.

M. SAIBANTE, C. VIVARINI, G. VOGHERA, *Gli studenti dell'Università di Padova dalla fine del 500 ai nostri giorni*, in « Metron », vol. IV, n. 1.

N. SPANO, *L'Università di Roma*, Roma, Casa ed. « Mediterranea ». 1935.

R. VALENTINI, *Studium Urbis*, in « Archivio della R. Deputazione Romana di Storia Patria », vol. II, nuova serie, 1936.

TABELLA I.

Studenti distinti per facoltà dall'anno 1753-54 all'anno 1773-74.

ANNI ACCADEMICI	Giurisprudenza	Medicina	Teologia	Filosofia e Matematica	Lingue	TOTALE
1753–54	105	12	16	4	1	138
1754–55	71	11	7	—	1	90
1755–56	134	21	13	8	1	177
1756–57	131	18	14	3	—	166
1757–58	180	21	13	8	1	223
1758–59	182	15	15	5	4	221
1759–60	108	15	15	8	3	149
1760–61	185	29	21	8	1	244
1761–62	143	23	24	4	1	195
1762–63	127	18	22	10	—	177
1763–64	123	22	19	9	3	176
1764–65	160	25	23	10	1	219
1765–66	164	23	35	6	1	229
1766–67	103	13	21	3	—	140
1767–68	149	14	9	10	1	183
1768–69	177	25	17	10	1	230
1769–70	158	21	19	10	3	211
1770–71	162	17	12	5	—	196
1771–72	150	21	9	7	1	188
1772–73	105	13	18	3	1	140
1773–74	151	14	21	8	1	195

TABELLA II.

Matricole distinte per facoltà dall'anno 1788-89 all'anno 1792-93.

ANNI ACCADEMICI	Giurisprudenza	Materie sacre	Medicina e Chirurgia	Filosofia ed Arti	Lingue	TOTALE
1788–89	124	18	76	11	14	243
1789–90	108	22	82	37	13	262
1790–91	103	17	72	37	8	237
1791–92	113	30	71	59	12	285
1792–93	86	22	76	31	18	233

TABELLA V.

Studenti italiani e stranieri classificati secondo la provenienza dall'anno 1753-54 all'anno 1773-74.

ANNI ACCADEMICI	Italia Settentr.	Italia Centrale*	*Lazio*	Italia Meridion.	Italia Insulare	Corsica	TOTALE ITALIANI	Africa	America	Asia Minore	Austria	Belgio	Boemia	Croazia	Francia	Germania	Grecia	Irlanda	Lussemburgo	Messico	Moravia	Polonia	Portogallo	Rep. S. Marino	Princ.° Monaco	Spagna	Svizzera	Transilvania	Ungheria	TOTALE STRANIERI	Provenienza ignota	TOTALE
1753–54	5	88	*72*	8	2	6	109	—	—	1	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	2	—	—	—	4	25	138
1754–55	10	50	*44*	5	2	6	73	—	—	1	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	1	3	14	90
1755–56	12	99	*84*	10	2	10	133	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	—	—	—	3	41	177
1756–57	15	97	*88*	10	1	8	131	—	—	2	—	—	—	—	—	—	1	—	—	—	—	1	1	—	—	—	—	—	—	5	30	166
1757–58	19	122	*103*	19	—	18	178	—	—	1	—	—	1	—	2	4	—	—	1	—	—	2	—	—	—	—	—	—	—	11	34	223
1758–59	16	129	*114*	12	1	14	172	—	—	2	1	—	1	—	1	8	—	1	—	—	—	—	—	—	—	—	1	—	—	15	34	221
1759–60	7	96	*86*	4	—	10	117	—	—	—	1	—	—	—	1	5	—	—	—	—	—	2	3	—	—	—	—	—	—	12	20	149
1760–61	16	151	*136*	14	1	13	195	—	—	2	—	—	1	—	3	14	—	—	1	—	—	1	—	—	—	—	—	—	—	22	27	244
1761–62	15	118	*104*	8	1	17	159	—	—	—	1	—	—	—	1	14	1	—	—	—	—	—	—	—	—	1	—	—	—	18	18	195
1762–63	11	112	*96*	10	3	10	146	—	—	—	—	—	—	—	—	15	—	—	—	—	1	3	—	—	—	—	—	—	—	19	12	177
1763–64	12	101	*88*	8	1	8	130	—	—	—	—	—	—	—	—	21	—	—	—	—	—	2	1	—	—	2	—	—	—	26	20	176
1764–65	20	119	*112*	17	1	14	171	—	—	—	—	1	—	—	5	15	—	—	—	—	1	—	1	—	—	1	—	1	—	25	23	219
1765–66	8	149	*125*	10	1	7	175	—	—	—	1	—	—	—	1	11	—	—	—	—	1	3	1	—	—	2	—	—	—	20	34	229
1766–67	7	101	*92*	7	1	8	124	—	—	—	—	—	—	—	—	3	—	—	—	—	—	1	—	—	—	2	—	—	1	7	9	140
1767–68	9	133	*115*	7	1	14	164	—	—	—	—	—	—	—	—	2	—	—	—	—	—	—	—	1	—	7	—	—	—	10	9	183
1768–69	14	146	*126*	7	2	16	185	—	—	1	—	—	—	1	1	11	—	—	—	—	—	2	—	1	1	11	—	—	—	29	16	230
1769–70	10	135	*120*	10	2	10	167	1	—	—	—	—	1	1	—	10	2	—	—	—	—	1	—	—	—	11	—	—	—	27	17	211
1770–71	10	137	*128*	6	2	15	170	—	—	—	—	1	—	—	—	10	3	—	—	1	—	1	—	—	—	2	—	—	—	18	8	196
1771–72	3	114	*99*	9	—	7	133	—	—	—	—	—	—	1	3	19	1	—	—	—	—	—	1	—	—	5	—	—	—	30	25	188
1772–73	4	85	*76*	2	—	12	103	—	1	—	—	—	—	—	1	7	—	—	—	—	—	—	—	—	—	—	—	—	—	9	28	140
1773–74	14	110	*97*	12	2	18	156	—	—	1	—	1	—	—	—	7	2	—	—	—	—	3	—	—	—	1	—	—	—	15	14	185

* Compreso il Lazio.

TABELLA III.

Percentuali degli studenti delle varie facoltà sul totale degli iscritti dall'anno 1753-54 all'anno 1773-74.

ANNI ACCADEMICI	Giurisprudenza	Medicina	Teologia	Filosofia e Matematica	Lingue	TOTALE
1753-54	76,1	8,7	11,6	2,9	0,7	100
1754-55	78,9	12,2	7,8	—	1,1	»
1755-56	75,7	11,9	7,3	4,5	0,6	»
1756-57	78,9	10,9	8,4	1,8	—	»
1757-58	80,7	9,4	5,8	3,6	0,5	»
1758-59	82,3	6,8	6,8	2,3	1,8	»
1759-60	72,5	10,1	10,1	5,3	2,0	»
1760-61	75,8	11,9	8,6	3,3	0,4	»
1761-62	73,3	11,8	12,3	2,1	0,5	»
1762-63	71,8	10,2	12,4	5,6	—	»
1763-64	69,9	12,5	10,8	5,1	1,7	»
1764-65	73,0	11,4	10,5	4,6	0,5	»
1765-66	71,6	10,1	15,3	2,6	0,4	»
1766-67	73,6	9,3	15,0	2,1	—	»
1767-68	81,4	7,7	4,9	5,5	0,5	»
1768-69	77,0	10,9	7,4	4,3	0,4	»
1769-70	74,9	10,0	9,0	4,7	1,4	»
1770-71	82,7	8,7	6,1	2,5	—	»
1771-72	79,8	11,2	4,8	3,7	0,5	»
1772-73	75,0	9,3	12,9	2,1	0,7	»
1773-74	77,4	7,2	10,8	4,1	0,5	»

TABELLA IV.

Percentuali delle matricole delle varie facoltà sul totale delle matricole dall'anno 1788-89 all'anno 1792-93.

ANNI ACCADEMICI	Giurisprudenza	Materie sacre	Medicina e Chirurgia	Filosofia ed Arti	Lingue	TOTALE
1788-89	51,6	7,4	31,3	4,5	5,8	100
1789-90	41,2	8,4	31,3	14,1	5,0	»
1790-91	43,4	7,2	30,4	15,6	3,4	»
1791-92	39,7	10,5	24,9	20,7	4,2	»
1792-93	36,9	9,5	32,6	13,3	7,7	»

TABELLA VI.

Percentuali degli studenti classificati secondo la provenienza dall'anno 1753-54 all'anno 1773-74.

ANNI ACCADEMICI	% Italiani	% Stranieri	% Provenienza ignota	TOTALE	% Italia Settentr.	% Italia Centrale *	% *Lazio*	% Italia Meridion.	% Italia Insulare	% Corsica	TOTALE ITALIANI
1753–54	79,0	2,9	18,1	100	4,6	80,7	*66,1*	7,4	1,8	5,5	100
1754–55	81,1	3,3	15,6	»	13,7	68,5	*60,3*	6,9	2,7	8,2	»
1755–56	75,1	1,7	23,2	»	9,0	74,5	*63,2*	7,5	1,5	7,5	»
1756–57	78,9	3,0	18,1	»	11,5	74,0	*67,2*	7,6	0,8	6,1	»
1757–58	79,9	4,9	15,2	»	10,7	68,5	*57,9*	10,7	—	10,1	»
1758–59	77,8	6,8	15,4	»	9,3	75,0	*66,3*	7,0	0,6	8,1	»
1759–60	78,5	8,1	13,4	»	6,0	82,1	*73,5*	3,4	—	8,5	»
1760–61	79,9	9,0	11,1	»	8,2	77,4	*69,7*	7,2	0,5	6,7	»
1761–62	81,6	9,2	9,2	»	9,4	74,3	*65,4*	5,0	0,6	10,7	»
1762–63	82,5	10,7	6,8	»	7,6	76,7	*65,8*	6,8	2,1	6,8	»
1763–64	73,9	14,8	11,3	»	9,2	77,7	*76,9*	6,2	0,7	6,2	»
1764–65	78,1	11,4	10,5	»	11,7	69,6	*65,5*	9,9	0,6	8,2	»
1765–66	76,4	8,8	14,8	»	4,6	85,1	*71,4*	5,7	0,6	4,0	»
1766–67	88,6	5,0	6,4	»	5,6	81,5	*74,2*	5,6	0,8	6,5	»
1767–68	89,6	5,5	4,9	»	5,5	81,1	*70,1*	4,3	0,6	8,5	»
1768–69	80,5	12,6	6,9	»	7,6	78,9	*68,1*	3,8	1,1	8,6	»
1769–70	79,1	12,8	8,1	»	6,0	80,8	*71,9*	6,0	1,2	6,0	»
1770–71	86,7	9,2	4,1	»	5,9	80,6	*75,3*	3,5	1,2	8,8	»
1771–72	70,7	16,0	13,3	»	2,3	85,7	*74,4*	6,8	—	5,2	»
1772–73	73,6	6,4	20,0	»	3,9	82,5	*73,8*	1,9	—	11,7	»
1773–74	84,3	8,1	7,6	»	9,0	70,5	*62,2*	7,7	1,3	11,5	»

TABELLA VII.

Matricole italiane e straniere classificate secondo la provenienza dall'anno 1788-89 all'anno 1792-93.

ANNI ACCADEMICI	Italia Settentr.	Italia Centrale *	*Lazio*	Italia Meridion.	Italia Insulare	Corsica	TOTALE ITALIANI	Asia Minore	Brasile	Costantinopoli	Ebrei	Francia	Germania	Malta	Nizza	Polonia	Sclo	Spagna	Svizzera	TOTALE STRANIERI	Provenienza ignota	TOTALE
1788–89	29	148	*131*	15	2	25	219	1	—	—	—	—	5	—	—	—	—	1	—	7	17	243
1789–90	31	183	*160*	6	1	16	237	—	1	—	—	2	4	—	—	—	—	—	1	8	23	268
1790–91	22	172	*150*	3	1	18	216	—	1	—	—	—	—	—	1	—	—	1	1	4	17	237
1791–92	41	192	*167*	9	—	16	258	—	—	—	—	—	5	1	1	2	2	1	—	12	15	285
1792–93	36	147	*126*	10	1	12	206	—	—	2	1	1	2	—	2	—	1	6	2	17	10	233

* V. nota a Tab. V.

TABELLA VIII.

Percentuali delle matricole classificate secondo la provenienza dall'anno 1788-89 all'anno 1792-93.

ANNI ACCADEMICI	% Italiani	% Stranieri	% Provenienza ignota	TOTALE	% Italia Settentr.	% Italia Centrale*	% *Lazio*	% Italia Meridion.	% Italia Insulare	% Corsica	TOTALE ITALIANI
1788–89	90,1	2,9	7,0	100	13,2	67,7	*59,8*	6,8	0,9	11,4	100
1789–90	88,4	3,0	8,6	»	13,1	77,2	*67,5*	2,5	0,4	6,8	»
1790–91	91,1	1,7	7,2	»	10,2	79,6	*69,4*	1,4	0,5	8,3	»
1791–92	90,5	4,2	5,3	»	15,9	74,4	*64,7*	3,5	—	6,2	»
1792–93	88,4	7,3	4,3	»	17,4	71,4	*61,2*	4,9	0,5	5,8	»

TABELLA IX.

Studenti italiani dall'anno 1753-54 all'anno 1773-74 classificati secondo la popolosità dei Comuni di provenienza.

ANNI ACCADEMICI	Comuni con oltre 50.000 abitanti (compresa Roma)	Comuni con oltre 50.000 abitanti (esclusa Roma)	Comuni con 20–50.000 abitanti	Comuni con 10–20.000 abitanti	Comuni con meno di 10.000 abitanti	Non classificabili * *	TOTALE
1753–54	35	3	6	5	53	4	103
1754–55	29	2	4	8	19	7	67
1755–56	62	8	12	7	32	10	123
1756–57	40	6	7	12	53	11	123
1757–58	75	11	10	12	51	12	160
1758–59	80	11	4	10	63	1	158
1759–60	43	4	1	8	52	3	107
1760–61	99	6	6	5	66	6	182
1761–62	69	6	11	2	57	3	142
1762–63	50	1	7	5	63	11	136
1763–64	58	3	7	5	47	5	122
1764–65	87	13	7	7	50	6	157
1765–66	82	9	6	9	66	5	168
1766–67	69	9	2	7	34	4	116
1767–68	74	5	2	11	63	—	150
1768–69	89	8	5	9	64	2	169
1769–70	92	10	3	7	55	—	157
1770–71	90	6	3	5	54	3	155
1771–72	64	5	5	7	50	—	126
1772–73	57	3	3	4	24	3	91
1773–74	58	6	3	9	66	2	138

(*) V. nota a tab. V.

(**) Abbiamo compreso qui, oltre gli studenti la cui provenienza non risultava, anche quelli per i quali era indicata la regione ma non il Comune.

TABELLA X.

Percentuali degli studenti italiani dall'anno 1753-54 all'anno 1773-74 classificati secondo la popolosità dei Comuni di provenienza.

ANNI ACCADEMICI	Comuni con oltre 50.000 abitanti (compresa Roma)	Comuni con oltre 50.000 abitanti (esclusa Roma	Comuni con abitanti 20–50.000	Comuni con abitanti 10–20.000	Comuni con meno di 10.000 abitanti	Non classificabili *	TOTALE
1753–54	34,0	2,9	5,8	4,9	51,4	3,9	100
1754–55	43,3	3,0	6,0	11,9	28,4	10,4	»
1755–56	50,4	6,5	9,8	25,7	26,0	8,1	»
1756–57	32,5	3,8	5,7	9,8	43,1	8,9	»
1757–58	46,9	7,0	6,2	7,5	31,9	7,5	»
1758–59	50,6	10,2	2,6	6,3	39,9	0,6	»
1759–60	40,2	3,7	0,9	7,5	48,6	2,8	»
1760–61	54,4	3,3	3,3	2,7	36,3	3,3	»
1761–62	48,6	4,2	7,7	1,4	40,2	2,1	»
1762–63	36,8	0,7	5,1	3,7	46,3	8,1	»
1763–64	47,6	2,5	5,7	4,1	38,5	4,1	»
1764–65	55,4	8,3	4,5	4,5	31,8	3,8	»
1765–66	48,8	5,4	3,6	5,3	39,3	3,0	»
1766–67	59,5	7,8	1,7	6,0	29,3	3,5	»
1767–68	49,4	3,3	1,3	7,3	42,0	—	»
1768–69	52,7	4,7	3,0	5,3	37,9	1,1	»
1769–70	58,6	6,4	1,9	4,5	35,0	—	»
1770–71	58,1	3,9	1,9	3,2	34,9	1,9	»
1771–72	50,8	4,0	4,0	5,5	39,7	—	»
1772–73	62,6	3,3	3,3	4,4	26,4	3,3	»
1773–74	42,0	5,0	2,2	6,5	47,8	1,5	»

* V. nota a Tab. IX.

TABELLA XI.

Laurec e altri gradi accademici distinti per facoltà dal 1827 al 1841.

ANNI	GIURISPRUDENZA*				MEDICINA E CHIRURGIA**				TEOLOGIA				FILOSOFIA E MATEMATICA				TOTALE			
	Lauree comuni	Lauree d'onore	Licenze	Baccellierati	Lauree comuni	Lauree d'onore	Licenze	Baccellierati	Lauree comuni	Lauree d'onore	Licenze	Baccellierati	Lauree comuni	Lauree d'onore	Licenze	Baccellierati	Lauree comuni	Lauree d'onore	Licenze	Baccellierati
1827	30	—	—	—	19	—	—	—									49	—	—	—
1828	34	—	—	—	47	—	—	—									81	—	—	—
1829	30	—	—	—	48	—	—	—									78	—	—	—
1830	36	—	—	—	52	—	—	—									88	—	—	—
1831	35	—	—	—	47	—	—	—									82	—	—	—
1832	37	—	—	—	36	—	—	—									73	—	—	—
1833	*mancano i dati*				46	—	11	3	13	—	—	—	7	—	2	4	66	—	13	7
1834					47	6	84	84	30	—	—	—	10	—	1	1	87	6	85	85
1835	49	8	56	78	55	4	48	61	30	—	—	—	3	1	—	146	137	13	104	285
1836	21	6	63	110	39	3	53	48	23	—	—	—	4	6	1	126	87	15	117	284
1837	22	5	58	92	21	3	50	43	17	—	—	—	6	1	—	20	66	9	108	155
1838	41	6	67	94	63	2	41	49	25	—	—	—	6	2	11	210	135	10	119	353
1839	33	5	64	91	30	4	50	49	19	—	—	—	7	5	11	115	89	14	125	255
1840	68	9	44	77	41	2	18	66	10	—	—	—	8	4	9	72	127	15	71	215
1841	*m. d.*	*m. d.*	*m. d.*	*m. d.*	*m. d.*	*m. d.*	*m. d.*	*m. d.*	10	—	—	—	2	—	—	83	12	—	—	83

* Risultano inoltre i seguenti diplomi di *Perizia di notaro*: 3 nel 1838 : 4 nel 1839 ; 5 nel 1840.
** Tra le licenze sono stati inclusi anche i diplomi di baccellierato e licenza ; tra le lauree, anche quelli di baccellierato, licenza e laurea.

PROF. CORRADO GINI, *Direttore responsabile.*

Roma — Tipografia Ippolito Failli.