CHAPTER XII

INTRODUCTION TO THE THEORY OF STATISTICAL FUNCTIONS

A. Differentiable Statistical Functions (Sections 1 and 2)

1. Statistical Functions.[1] Continuity, Differentiability

1.1. *Distributions.* A distribution is denoted by $V(x)$ if it may be either a probability distribution $P(x)$ or a sample distribution $S_n(x)$ ("repartition") of n results. We review the definition of $S_n(x)$. Given n real numbers $x_1, x_2, \ldots, x_n$, the number of those x_ν ($\nu = 1, 2, \ldots, n$) which are less than or equal to x is denoted by $n\,S_n(x)$. Consequently, $S_n(x)$ is a stepline or step function increasing from 0 to 1 whose steps are at the points x_ν with heights equal to multiples of $1/n$. We assume that each x_ν is the result of an observation which is subject to a probability distribution $P_\nu(x)$, $\nu = 1, 2, \ldots, n$ and that $P_\nu(x)$ is either absolutely continuous or a step function, or a combination of both.

We consider, in the following, sets J of distributions. J may also be regarded as a certain subspace, a region of the *space of distributions* where each $V(x)$ is a "point." Let $V_1(x)$ and $V_2(x)$ be two "points" of J;

[1] In the last section of Chapter XI we introduced the term statistical function and added some comments. The papers by R. v. Mises on which the present chapter is based will be quoted as A, B, C, D, E, respectively: they are:

(A) "Deux nouveaux théorèmes de limite dans le calcul des probabilités." *Rev. Fac. Sci. Univ. Istanbul* 1 (1935), pp. 61-80.

(B) "Die Gesetze der grossen Zahl für statistische Funktionen." *Monatsh. Math. Phys.* **43** (1936), pp. 105-128.

(C) "Les lois de probabilité pour les fonctions statistiques." *Ann. Inst. H. Poincaré* **6** (1936), pp. 185-212.

(D) "On the asymptotic distribution of differentiable statistical functions." *Ann. Math. Statist.* **18** (1947), pp. 309-348.

(E) "Théorie et applications des fonctions statistiques." *Rend. Mat. appl.* **11** (1953), pp. 1-37.

they define the "straight segment" from $V_1(x)$ to $V_2(x)$, which contains all distributions of the form

$$V_1(x) + t[V_2(x) - V_1(x)], \qquad 0 \leqslant t \leqslant 1. \tag{1}$$

We call J *convex* if, together with two distributions $V_1(x)$ and $V_2(x)$, it contains all distributions of the form (1); in other words, if it includes the segment determined by any two of its points.

1.2. *Statistical functions. Functionals.* Let J be a convex set of distributions $V(x)$. To each $V(x)$ of J we associate a number f and call $f\{V(x)\}$ (or $f\{P(x)\}$, or $f\{S_n(x)\}$) *a statistical function* on J. If $V(x) = S_n(x)$, then $f\{S_n(x)\}$ is simply a function of the n variables $x_1, x_2, \ldots, x_n$, which has the following properties: (1) It is *symmetric* in the $x_1, x_2, \ldots, x_n$; that means the value of f does not change if any x_κ is interchanged with any x_λ. (2) The value of f does not change if n is replaced by $2n$, $3n$, ... and each x_ν ($\nu = 1, 2, \ldots, n$) is counted twice, three times, ... among the $2n$, $3n$, ... variables. If in $f\{V(x)\}$ the $V(x)$ is a discrete distribution with probabilities (or relative frequencies) $p_1, p_2, \ldots, p_k$, then $f\{V(x)\}$ is simply a function of the k numbers $p_1, p_2, \ldots, p_k$. This is a situation considered in Chapter VII, Section 6 and at the end of Chapter XI.[2] In the case of a discrete distribution with *countably* many attributes (Chapter I) $f\{V(x)\}$ is a function of the countably many numbers p_1, $p_2, \ldots$. We may either generalize the considerations and results which hold for arithmetical $V(x)$ with k attributes or treat $f\{V(x)\}$ as a general statistical function.

While in these special cases $f(V(x)\}$ reduces to a function of several or countably many variables, it is in general a functional[3] $f\{V(x)\}$ or,

[2] In Chapter XI we could not write $r_1, r_2, \ldots, r_k$ for the relative frequencies (as done in Chapters VII and X and others) since in XI, the r was identified with various measures of correlation; hence we used p_i for the frequencies and π_i for the probabilities. Here now, we use again, as in the greater part of the book, p_i and $p(x)$ for probabilities and r_i for frequencies.

[3] Such functionals have been introduced foremost by V. Volterra, but also by M. Fréchet, J. Hadamard, and many other mathematicians. We quote V. Volterra:

(a) "Sopra le funzioni che dipendonno da altre funzioni." *Rend. Accad. Lincei* **6** (1887); three notes: pp. 97, 141, 153.

(b) "Sopra le funzioni dipendente da linee." *ibid.* pp. 225 and 274.

(c) *Leçons sur les fonctions de ligne.* Paris, 1913.

(d) Vito Volterra et Joseph Pérès, *Théorie générale des fonctionnelles.* Paris, 1936.

(e) *Theory of Functionals and of Integral-and Integro-Differential Equations.* New York, 1959.

The name "functional" is due to Hadamard and has been fairly generally accepted while Volterra's "fonction de ligne" = "line function" is less used.

more specifically, *a functional of* $V(x)$. A quantity $z = f\{y(\overset{b}{\underset{a}{x}})\}$ is a functional of a function $y(x)$ in $a \leqslant x \leqslant b$ when a law is given which puts every function $y(x)$ of a certain type in correspondence with a quantity (number) $z = f\{y(\overset{b}{\underset{a}{x}})\}$. If $a = -\infty$, $b = +\infty$, we omit these limits in the notation. If the independent variable $y(x)$ is a distribution the functional $f\{V(x)\}$ is called a *statistical function*.

To visualize the general case where $f\{y(x)\}$ does not reduce to an ordinary function of several variables, we start with a polygonal line inscribed in an arc of a curve. (Fig. 52.) The vertices of the polygon have abscissas

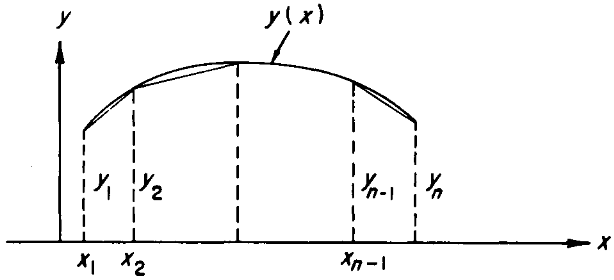

FIG. 52. Curve with inscribed polygonal line.

x_i and ordinates y_i. Then, $F(y_1, y_2, \ldots, y_n)$ is a function of the polygonal line. Any curve can be considered as the limit of inscribed polygons with ever increasing number n of sides, when, at the same time, these sides decrease in length. In the limit, a continuous parameter takes the place of the discontinuous subscript i and sums with respect to i are replaced by integrals. Following Volterra[4] we shall regard $f\{y(x)\}$ as the limit of $F(y_1, y_2, \ldots, y_n)$.

The simplest example of a statistical function is the mean $\int x\, dV(x)$ or, more generally, the *linear statistical function*

$$f_1\{V(x)\} = \int a(x)\, dV(x), \tag{2}$$

where $a(x)$ is a given continuous function of x only, not of $V(x)$, and $V(x)$ a (variable) distribution for which the Stieltjes integral on the right-hand side converges. Clearly, (2) generalizes the linear form in n variables $\sum_{i=1}^{n} a_i y_i$. The discontinuous i is replaced by x and the sum

[4] We shall use this analog as a guide, a help for better understanding of the definitions. Volterra, however, derives the formulas and definitions for functionals from those for $F(y_1, y_2, \ldots, y_n)$ by means of a passage to the limit.

by an integral. We know (Chapter II) that $V(x)$ is, most generally, the sum of an absolutely continuous distribution $V(x) = \int^x (dV/dt)\, dt$, and a discrete distribution with at most countably many discontinuities of the first kind [the "singular" part of a monotonic $V(x)$ (Chapter II, p. 96) does not appear in our frequency theory of probability]. Therefore, with $v(x)$ for the density and v_i for the heights of the jumps the general linear function which we consider is of the form

$$\int a(x)v(x)\, dx + \sum_i a(x_i)v_i\,. \tag{2'}$$

If we introduce in (2): $V(x) = \lambda_1 V_1(x) + \lambda_2 V_2(x)$, $\lambda_1 + \lambda_2 = 1$, we obtain

$$\int a(x)\, dV(x) = \lambda_1 \int a(x)\, dV_1(x) + \lambda_2 \int a(x)\, dV_2(x),$$

the characteristic relation of linearity. If $V(x) = S_n(x)$ where $S_n(x)$ is the repartition (sample distribution) of $x_1, x_2, \ldots, x_n$, then

$$\int a(x)\, dS_n(x) = \frac{1}{n}[a(x_1) + a(x_2) + \cdots + a(x_n)]. \tag{3}$$

Next, we consider the *bilinear* statistical function defined by means of a given continuous and symmetric $b(x, y)$:

$$f_2\{V(x)\} = \iint b(x, y)\, dV(x)\, dV(y). \tag{4}$$

This corresponds to $\sum_{i=1}^n \sum_{j=1}^n b_{ij}y_iy_j$. We can always assume that $b_{ij} = b_{ji}$, for, if not, we put, $\sum_{i,j} b_{ij}y_iy_j = \sum_{i,j} b'_{ij}y_iy_j$ where $b'_{ij} = \frac{1}{2}(b_{ij} + b_{ji})$; the same holds for $b(x, y) = b(y, x)$ in (4). More generally, a *homogeneous statistical function of degree r* is given by

$$f_r\{V(x)\} = \int \cdots \int k(x_1, x_2, \ldots, x_r)\, dV(x_1)\, dV(x_2) \cdots dV(x_r), \tag{5}$$

where $k(x_1, x_2, \ldots, x_r)$ is symmetric.

As important examples of non-linear statistical functions we mention the moments with respect to the mean

$$\begin{aligned} f = m_r &= \int (x - a)^r\, dS_n(x), \qquad a = \int x\, dS_n(x) \\ &\text{or} \\ \mu_r &= \int (x - \alpha)^r\, dP(x), \qquad \alpha = \int x\, dP(x), \end{aligned} \tag{6}$$

or the Lexis quotient

$$L = \frac{Nm_2}{a(N-a)} \tag{7}$$

(N constant). Student's $t = \sqrt{n-1}\,\frac{\bar{x}-\alpha}{s}$, Pearson's X^2, v. Mises' or Smirnov's ω^2,

$$\omega^2 = \int \lambda(x)[S_n(x) - P(x)]^2\, dx; \tag{8}$$

the median, the quartiles, etc., are further examples. Another example of a statistical function is the greatest of n sample values. Observations are made from the same collective with distribution $P(x)$; X^0 the greatest of the first n sample values, is a statistical function which is neither continuous nor "differentiable" (in a sense which we shall study presently) and to which our theory does not apply.

The concept of a statistical function applies to collectives of more than one dimension. For example, the correlation coefficient is a statistical function.

1.3. *Continuity.* Consider a set J of distributions [i.e., a region J in the space of distributions where each $V(x)$ is a point]. We call $V(x)$ a *bounded distribution* if there exist two finite numbers A and B ($A < B$) such that $V(A) = 0$, $V(B) = 1$. An infinite set $V_\nu(x)$ of bounded distributions is called *uniformly bounded* if bounds A and B can be found independent of ν. Let us now assume that J is a uniformly bounded set of distributions. A function $f\{V(x)\}$ is called *continuous* at $V_1(x)$ if for $\epsilon > 0$ we can find an η such that, if $V(x)$ is any distribution in J

$$|f\{V(x)\} - f\{V_1(x)\}| \leqslant \epsilon \qquad \text{if} \qquad |V(x) - V_1(x)| \leqslant \eta \tag{9}$$

for all x in the range under consideration. Thus, continuity of $f\{V(x)\}$ means: If the variation of $V_1(x)$ is small the statistical function $f\{V_1(x)\}$ changes little.

If, *in the general case of unbounded distributions* we were to ask for this strong continuity, the concept would be too narrow. Consider, for example, the linear case $f = \int a(x)dV(x)$. This continuity would not even hold for moments, where $a(x) = x^m$. Hence, let $\psi(x)$ *be a positive, continuous function which increases with* $|x|$. We call $f\{V(x)\}$ *continuous at* $V_1(x)$ *with respect to* $\psi(x)$ if, for all $V(x)$ of J, an η exists to a given ϵ, such that

$$|f\{V(x)\} - f\{V_1(x)\}| \leqslant \epsilon \qquad \text{if} \qquad \psi(x)\,|V(x) - V_1(x)| \leqslant \eta. \tag{9'}$$

For bounded distributions continuity becomes independent of $\psi(x)$.[5] We see, for example, that a linear statistical function is continuous for a $V(x)$ for which it exists with respect to any $\psi(x)$ for which $\int | a(x) | \, d(1/\psi)$ converges as $| x | \to \infty$. In the case of the mean, $\int x \, dV$, we may take $\psi(x) = | x |^{1+\rho}$, $\rho > 0$.

1.4. *Differentiability.* We come now to the important concept of differentiation. For the function $F(y_1, y_2, ..., y_n)$ we assume existence of the partial derivatives $\partial F/\partial y_i$ and a total differential $dF = \sum_{i=1}^{n} (\partial F/\partial y_i) dy_i$. The total differential dF has two well-known basic properties: (a) it is a linear form in the n differentials dy_i; (b) if we set $dy_i = \epsilon k_i$ and take ϵ small, then dF differs from $\Delta F = F(y_1 + dy_1, y_2 + dy_2, ..., y_n + dy_n) - F(y_1, y_2, ..., y_n)$ by a small quantity of higher order than ϵ. Indeed

$$\begin{aligned} F + \Delta F &= F(y_1 + dy_1, y_2 + dy_2, ..., y_n + dy_n) \\ &= F(y_1, y_2, ..., y_n) + \sum_{i=1}^{n} \frac{\partial F}{\partial y_i} dy_i + o(\epsilon) \qquad (10) \\ &= F + dF + o(\epsilon) . \end{aligned}$$

Hence,

$$\Delta F = dF + o(\epsilon). \qquad (10')$$

A function $F(y_1, y_2, ..., y_n)$ is called *differentiable* if a formula (10′) exists with $dF = \sum_{\nu=1}^{n} A_\nu dy_\nu$, the A_ν being independent of the dy_ν. We wish to define differentiation of functionals in analogy to the above concepts.[6]

Consider first $\partial F/\partial y_i$; it is defined as the limit, if it exists, of $\Delta_i F/\Delta y_i$ where Δy_i is an increment of y_i alone while all other y_j, $j \neq i$, remain unchanged, and $\Delta_i F$ is the increment of F corresponding to Δy_i. To obtain the analog of the partial derivative for $f\{y(x)\}$ at $x = \xi$, we vary $y(x)$ along a segment α, β of length h which contains the abscissa ξ, assuming that f is defined for the varied $y(x)$ (Fig. 53). Let $\delta y(x) = \phi(x)$ be such an increment which does not change sign between α and β, and let $| \phi(x) | < \epsilon$ in α, β and $\phi(x) = 0$ outside α, β. We assume $y(x)$ to be continuous and $f\{y(x)\}$ continuous in the sense of (9), and we shall make the following further assumptions:

(a) With the given meaning of ϵ and h, and $\Delta f = f\{y(x) + \phi(x)\} - f\{y(x)\}$, the ratio $\Delta f/\epsilon h$ is always less than a finite fixed number M,

$$\frac{\Delta f}{\epsilon h} < M.$$

[5] If $V(x)$ is bounded, we may replace $\psi(x)$ by $\psi_{\max}$ and η by $\psi_{\max} \cdot \eta$.

[6] Volterra, p. 24 of footnote 3(c), cited p. 616.

(b) Denote by σ the small area over α, β between $y(x)$ and $y(x) + \phi(x)$, where the sign of $\phi(x)$ does not change. It is assumed that the double limit

$$\lim_{\epsilon \to 0, h \to 0} \frac{\Delta f}{\sigma} \tag{11}$$

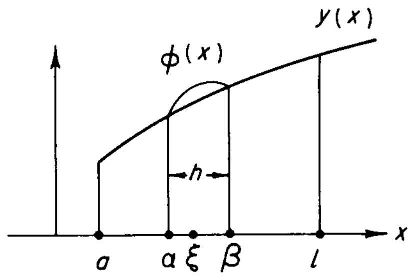

FIG. 53. Variation of a curve over small interval.

exists if simultaneously $\epsilon \to 0$, $h \to 0$ and ξ is always in (α, β); we assume that the limit is the same whether $\phi(x) \geqslant 0$ or $\leqslant 0$ in α, β.

(*c*) The ratio $\Delta f/\sigma$ tends to its limit uniformly for all points ξ in the interval a, b and for all $y(x)$.

The limit (11) which will depend in general, on $y(x)$ and on ξ is denoted by

$$f'\{y(x), \xi\} \tag{12}$$

(d) We suppose that $f'\{y(x), \xi\}$ is continuous with respect to $y(x)$ and to ξ.

We then call (12) *the first (partial) derivative of the functional f* with respect to $y(x)$ *at* ξ. Since ξ can take any value in an interval there exist infinitely many partial derivatives; the continuous parameter ξ takes the place of the subscript i in $\partial F/\partial y_i$.

To arrive at the concept which corresponds to the total differential we assume that the line $y(x)$ is deformed a little along its *total* length between a and b by $\delta y(x) = \epsilon k(x)$. We wish to compute that part of the increment Δf which corresponds to the dF in the case of $F(y_1, y_2, \ldots, y_n)$. (This will be the first variation of the functional f.) Turning back to (10) we remember the essential properties of dF, namely, that dF is linear in the n quantities $dy_i = \epsilon k_i$ and that $\Delta F - dF$ is small of a higher order than ϵ.

In the case of a functional $f\{y(x)\}$ the "first variation" δf which now takes the place of dF should be a linear function of $\delta y(x) = \epsilon k(x)$, and $\Delta f - \delta f$ should be $o(\epsilon)$. Volterra (footnotes 3(a) and 3(b), cited p. 616) has established a formula for δf which satisfies these two requirements, under the assumption that $f\{y(x)\}$ has a partial derivative $f'\{y(x), \xi\}$ continuous with respect to ξ and continuous with respect to $y(x)$. The sum on the right-hand-side in (10) is to be replaced by an integral, the $\partial F/\partial y_i$ by $f'\{y(x), \xi\}$ and the dy_i by $\delta y(\xi)$. The result is

$$\partial f = \int_a^b f'\{y(x), \xi\} \, \partial y(\xi) \, d\xi; \tag{13}$$

δf is called the *differential*, or first variation, of $f\{y(x)\}$.

We wish to rewrite (13) in another form. For this purpose we return once more to $F(y_1, y_2, ..., y_n)$, put $dy_i = \epsilon k_i$ and set $z_i = y_i + \epsilon k_i$, $i = 1, 2, ..., n$. Then

$$\frac{\Delta F}{\epsilon} = \frac{F(z_1, ..., z_n) - F(y_1, ..., y_n)}{\epsilon} = \sum_{i=1}^{n} \frac{\partial F}{\partial y_i} k_i + \frac{o(\epsilon)}{\epsilon}$$

and

$$\frac{d}{d\epsilon} F(z_1, ..., z_n) = \frac{\partial F}{\partial z_1} k_1 + \cdots + \frac{\partial F}{\partial z_n} k_n .$$

Letting $\epsilon \to 0$, we obtain

$$\lim_{\epsilon \to 0} \frac{\Delta F}{\epsilon} = \lim_{\epsilon \to 0} \frac{dF}{\epsilon} = \sum_{i=1}^{n} \frac{\partial F}{\partial y_i} k_i = \left[\frac{d}{d\epsilon} F(y_1 + \epsilon k_1, ..., y_n + \epsilon k_n)\right]_{\epsilon=0} . \tag{14}$$

In the same way, we have in the case of functionals, using the decisive fact that $\lim_{\epsilon\to 0} (\Delta f/\epsilon - \delta f/\epsilon) = 0$, the analogous equation

$$\lim_{\epsilon \to 0} \frac{\partial f}{\epsilon} = \int_a^b f'\{y(x), \xi\} k(\xi) \, d\xi = \left[\frac{d}{d\epsilon} f\{y(x) + \epsilon k(x)\}\right]_{\epsilon=0} . \tag{15}$$

Here $\epsilon k(\xi)$ takes the place of $\delta y(\xi)$ in Eq. (13). The last equality in (15) forms the analog of the definition of differentiability of a function of several variables: *If a relation*

$$\int_a^b f'\{y(x), \xi\} k(\xi) \, d\xi = \left[\frac{d}{d\epsilon} f\{y(x) + \epsilon k(x)\}\right]_{\epsilon=0} \tag{15'}$$

holds for $\delta y(x) = \epsilon k(x)$, $f\{y(x)\}$ *is called differentiable and the function* $f'\{y(x),\xi\}$ *so defined its* (first partial) *derivative at* ξ.

We turn to statistical functions.[7] For $y(x)$ we take the distribution

[7] The reader should note that the notation of v. Mises differs from that of Volterra. Take Eq. (2) which reads $f\{V(x)\} = \int a(x) \, dV(x)$. If $V(x)$ has a density $v(x)$, we have $\int a(x) \, dV(x) = \int a(x) \, v(x) \, dx$ and, in Volterra's notation, this would be $f\{v(x)\}$. It should be easy to avoid confusion.

$V_1(x)$, for the increment $k(\xi)\,d\xi$ we take $dV(\xi) - dV_1(\xi)$ which we shall often write as $d(V - V_1)(\xi)$. Put $W(x) = V_1(x) + t[V(x) - V_1(x)]$; then $f\{W(x)\}$ is an ordinary function of t and we assume that it has derivatives with respect to t up to a certain order, at least around $t = 0$. The condition (15′) takes then the form

$$\left[\frac{d}{dt} f\{V_1(x) + t[V(x) - V_1(x)]\}\right]_{t=0} = \int_a^b f'\{V_1(x), \xi\}\, d(V - V_1)(\xi), \quad (16)$$

where the integral in (16) exists. *If a relation* (16) *holds for any two "points"* $V(x)$, $V_1(x)$ *of the convex region* J *under consideration,* f *is called differentiable at the point* $V_1(x)$. The first derivative depends on $V_1(x)$ and on a scalar variable ξ, $-\infty < \xi < \infty$. We note that by (16), an "additive constant" in f', that is, a quantity not dependent on ξ, has no significance since the integral $\int d(V - V_1)(x)$ vanishes.

1.5. *Some examples.* Consider a few applications. The derivative of a linear statistical function $\int a(x)\, dV(x)$ equals $a(x)$, independent of ξ, as seen immediately from (16).

Consider (4). With $W = V_1 + t(V - V_1)$ we have

$$f_2\{W\} = \iint b(x, y)\, d[V_1(x) + t(V(x) - V_1(x))]\, d[V_1(y) + t(V(y) - V_1(y))],$$

$$\begin{aligned}\left[\frac{d}{dt} f_2\{W\}\right]_{t=0} &= \iint b(x, y)\, dV_1(x)\, d(V - V_1)(y) \\ &\quad + \iint b(x, y)\, dV_1(y)\, d(V - V_1)(x) \\ &= 2 \iint b(x, y)\, dV_1(x)\, d(V - V_1)(y).\end{aligned}$$

If we set according to (16)

$$2 \iint b(x, \xi)\, dV_1(x)\, d(V - V_1)(\xi) = \int f_2'\{V_1(x), \xi\}\, d(V - V_1)(\xi)$$

we see that the differentiation formula, with V instead of V_1 reads

$$f_2'\{V(x), \xi\} = 2 \int b(x, \xi)\, dV(x). \quad (4')$$

We have used the fact that $b(x, y)$ is symmetric. Otherwise we would have obtained

$$f_2'\{V(x), \xi\} = \int [b(x, \xi) + b(\xi, x)]\, dV(x). \quad (4'')$$

Next, consider Eq. (5). We could obtain, in a similar way, the derivative of $f_r\{V(x)\}$, $r = 2, 3, \ldots$. However, since we shall later need the Taylor expansion of $V(x)$ about an arbitrary point, we consider a generalized expression (5) where $V - V_0$ plays the role of V. We put $T(x) = V(x) - V_0(x)$ and consider with a symmetric $k(x_1, x_2, \ldots, x_r)$:

$$g_r\{V(x)\} = \int \cdots \int k(x_1, x_2, \ldots, x_r)\, dT(x_1)\, dT(x_2) \cdots dT(x_r), \tag{17}$$

where $g_r\{V_0(x)\} = 0$. Obviously, this is the analog of a nonhomogeneous polynomial of degree r. To find the characteristic properties of g_r, we wish to find the derivative of $g_r\{V(x)\}$ at an arbitrary point $V_1(x)$. To this purpose [in order to be able to use definition (16)] we must replace the $V(x)$ in (17) by the previously introduced $W(x) = V_1(x) + t[V(x) - V_1(x)]$; then $T(x) = V(x) - V_0(x)$ is replaced by $V_1(x) - V_0(x) + t[V(x) - V_1(x)]$. We then differentiate the product

$$\prod_{i=1}^{r} d[(V_1 - V_0)(x_i) + t(V - V_1)(x_i)]$$

with respect to t and set $t = 0$. This derivative consists of r terms, the first of which will be

$$d(V - V_1)(x_1) \cdot \prod_{i=2}^{r} d(V_1 - V_0)(x_i).$$

To find $[(d/dt)g_r\{V(x)\}]_{t=0}$ we have to multiply the sum of these r terms by $k(x_1, x_2, \ldots, x_r)$ and to integrate [as on the right-hand side of (17)]. Since k is symmetric, all r terms supply the same integral and the derivative of g_r with respect to t at $t = 0$ becomes (with an r-tuple integral)

$$r \int \cdots \int k(x_1, x_2, \ldots, x_r) \prod_{i=2}^{r} d(V_1 - V_0)(x_i)\, d(V - V_1)(x_1).$$

Comparing this with Eq. (16) which defines the derivative of a statistical function, and writing finally ξ instead of x_1, and $V(x)$ instead of $V_1(x)$ in the $\Pi_{i=2}^{r}\, d(V_1 - V_0)(x_i)$ we find (with an $(r - 1)$-tuple integral):

$$g_r'\{V(x), \xi\} = r \int \cdots \int k(\xi, x_2, x_3, \ldots, x_r)\, d(V - V_0)(x_2) \cdots d(V - V_0)(x_r), \tag{18}$$

the first derivative of the inhomogeneous polynomial (17) *at the point* $V(x)$.

Equation (18) *shows that* $g_r'\{V_0(x), \xi\} = 0$ *for any* ξ. Now a few more examples:

The derivative of a function of the type

$$f\{V(x)\} = \int \phi[V(x)]\, dx, \tag{19}$$

where $\phi[V(x)]$ is an ordinary function of V is given by

$$f'\{V(x), \xi\} = -\int^{\xi} \phi'[V(x)]\, dx, \tag{19'}$$

where ϕ' is the derivative of ϕ with respect to V and integration by parts has been used.

If f is an ordinary function of several functionals $f_1\{V(x)\}$, $f_2\{V(x)\}$, ...,

$$f\{V(x)\} = F(f_1, f_2, \ldots) \tag{20}$$

the derivative of f is

$$\frac{\partial F}{\partial f_1} f_1'\{V(x), \xi\} + \frac{\partial F}{\partial f_2} f_2'\{V(x), \xi\} + \cdots . \tag{20'}$$

If, in particular $f_1 \equiv A = \int a(x)\, dV(x)$, $f_2 \equiv B = \int b(x)\, dV(x)$, ...

$$f = F(A, B, \ldots),$$

we find directly, or applying (20′):

$$f'\{V(x), \xi\} = \frac{\partial F}{\partial A} a(\xi) + \frac{\partial F}{\partial B} b(\xi) + \cdots . \tag{20''}$$

We compute the derivative of the central moment m_ν given in (6). Using (20′) we obtain

$$\begin{aligned} f'\{V(x), \xi\} &= (\xi - a)^\nu - \nu\xi \int (x - a)^{\nu-1}\, dV(x) \\ &= (\xi - a)^\nu - \xi\nu m_{\nu-1} . \end{aligned} \tag{6'}$$

The ξ in the second term on the right-hand side comes from the differentiation of $a = \int x\, dV(x)$.[8]

[8] Compare the computation (done in Chapter X) in the case of an arithmetical distribution: with $n_\kappa/n = r_\kappa$ we have $f = \sum_{\kappa=1}^{k} (x_\kappa - a)^\nu r_\kappa$, $a = \sum x_\kappa r_\kappa$. Then $\partial f/\partial r_j = -\sum_{\kappa=1}^{k} \nu(a_\kappa - a)^{\nu-1} r_\kappa \cdot x_j + (x_j - a)^\nu = -\nu m_{\nu-1} x_j + (x_j - a)^\nu$. Clearly, the x_j becomes ξ in the continuous case.

We consider finally the derivative of

$$\omega^2 = \int \lambda(x)[S_n(x) - P(x)]^2\,dx = f\{S_n(x)\}. \tag{21}$$

Replacing $S_n(x)$ by $V_1(x) + t[V(x) - V_1(x)] = W(x)$ and differentiating with respect to t we have, as in (16):

$$\begin{aligned}\frac{d}{dt}f\{W(x)\}\Big|_{t=0} &= \int f'\{V_1(x), \xi\}\,d(V - V_1)(\xi)\\ &= 2\int \lambda(x)[V_1(x) - P(x)][V(x) - V_1(x)]\,dx.\end{aligned}$$

Partial integration gives the derivative at the point V_1:

$$f'\{V_1(x), \xi\} = -2\int^{\xi} \lambda(x)[V_1(x) - P(x)]\,dx. \tag{21'}$$

It is seen that this is identically zero for $P(x) = V_1(x)$.

2. *Higher Derivatives. Taylor's Theorem*

2.1. *Higher derivatives.* Consider $f'\{V(x), \xi\}$ for constant $\xi = \xi_1$; if this functional of $V(x)$ satisfies the conditions given on p. 621, one can compute its derivative at ξ_2. The result is a functional denoted by

$$f''\{V(x), \xi_1, \xi_2\}, \tag{22}$$

the *second derivative* of $f\{V(x)\}$. The two abscissas ξ_1, ξ_2 correspond to the two subscripts in $\partial^2 f/\partial y_i\,\partial y_j$. Volterra has proved (see Volterra, footnote 3(a), cit. p. 616) that under certain restrictions, a symmetry relation analogous to $\partial^2 f/\partial y_i\,\partial y_j = \partial^2 f/\partial y_j \partial y_i$ holds for (22):

$$f''\{V(x), \xi_1, \xi_2\} = f''\{V(x), \xi_2, \xi_1\}. \tag{22'}$$

Applying the definition (16) to $f'\{V(x), \xi_1\}$ we obtain the direct definition of a "twice differentiable" functional

$$\left[\frac{d^2}{dt^2}f\{V_1(x) + t[V(x) - V_1(x)]\}\right]_{t=0} = \iint f''\{V_1(x), \xi_1, \xi_2\}\,d(V - V_1)(\xi_1)\,d(V - V_1)(\xi_2) \tag{23}$$

or with $V - V_1 = T$ and writing $f''(\xi_1, \xi_2)$ for $f''\{V_1, \xi_1, \xi_2\}$

$$\left[\frac{d^2}{dt^2}f\{V_1(x) + tT(x)\}\right]_{t=0} = \iint f''(\xi_1, \xi_2)\,dT(\xi_1)\,dT(\xi_2). \tag{23'}$$

Note that the addition of a function of either ξ_1 or ξ_2 alone to $f''(\xi_1, \xi_2)$ does not change (23′), since the double integral of such an additive term with respect to $dT(\xi_1)\,dT(\xi_2)$ is zero.

We shall need in the following a certain "boundedness" of the second derivative which we define as follows[1]:

$$\left|\int\int f''(\xi_1, \xi_2)\,dT(\xi_1)\,dT(\xi_2)\right| \leqslant \int T^2(x)\psi(x)\,dx, \tag{24}$$

where $\psi(x) > 0$ is an appropriate function [this $\psi(x)$ has nothing to do with the $\psi(x)$ in (9′)].

As an application, consider again $f\{V(x)\} = F(\int a(x)\,dV, \int b(x)\,dV, \ldots)$; if we set, as before, $A = \int a(x)\,dV(x)$, $B = \int b(x)\,dV(x)$, ... we have

$$f''(\xi_1, \xi_2) = \frac{\partial^2 F}{\partial A^2} a(\xi_1)a(\xi_2) + 2\frac{\partial^2 F}{\partial A \partial B} a(\xi_1)b(\xi_2) + \frac{\partial^2 F}{\partial B^2} b(\xi_1)b(\xi_2) + \cdots.$$

The second derivative of a linear functional $\int a(x)dV(x)$ is zero since $a(x)$ does not depend on $V(x)$.

For a derivative of order r we have the definition [see (23′) and (16)] with $T = V - V_1$

$$\left[\frac{d^r}{dt^r} f\{V_1(x) + tT(x)\}\right]_{t=0} = \int \cdots \int f^{(r)}(\xi_1, \xi_2, \ldots, \xi_r)\,dT(\xi_1)\,dT(\xi_2)\cdots dT(\xi_r). \tag{25}$$

Here $f^{(r)}\{V_1(x), \xi_1, \ldots, \xi_r\} \equiv f^{(r)}(\xi_1, \ldots, \xi_r)$ is the derivative of $f^{(r-1)}\{V_1(x)$ $\xi_1, \ldots, \xi_{r-1}\}$, considered as a functional of V_1.

We note again that in expression (25) a term added to $f^{(r)}(\xi_1, \xi_2, \ldots, \xi_r)$ which does not depend on all r variables $\xi_1, \xi_2, \ldots, \xi_r$ can be added or discarded.

The rth derivative of the central moment m_ν is

$$f^{(r)}\{V(x), \xi_1, \xi_2, \ldots, \xi_r\}$$
$$= \frac{(-1)^r \nu!}{(\nu - r)!}\xi_1\xi_2\cdots\xi_r\left[m_{\nu-r} - \frac{1}{\nu - r + 1}\sum_{\rho=1}^{r}\frac{(\xi_\rho - a)^{\nu-r+1}}{\xi_\rho}\right]. \tag{6″}$$

Next we wish to compute the second derivative of the g_r in (17). Consider expression (17) and the result (18). The integral in (18) has the same form as that in (17) except that its multiplicity is $(r - 1)$ rather than r. Thus it is immediately seen how the higher derivatives of

[1] If the argument in $f'(V_i, \xi)$ or in $f''\{V_i, \xi_1, \xi_2\}$ is V_i then $T = V - V_i$ where $i = 0$ or $i = 1$.

(17) can be found. For $f_r''\{V(x), \xi_1, \xi_2\}$ we have to multiply (18) by $r - 1$, replace x_2 by ξ_2 in $k(\xi, x_2, ..., x_r)$ and omit in the product of differentials the $d(V - V_0)(x_2) = dT(x_2)$. Then

$$g_r''\{V(x), \xi_1, \xi_2\} = r(r-1) \int \cdots \int k(\xi_1, \xi_2, x_3, ..., x_r)\, dT(x_3) \cdots dT(x_r) \tag{18'}$$

and the rth derivative will be

$$g_r^{(r)}\{V(x), \xi_1, \xi_2, ..., \xi_r\} = r!k(\xi_1, \xi_2, ..., \xi_r), \tag{18''}$$

independent of $V(x)$, and the $(r + 1)$th derivative of (17) is zero. We note from (18), (18'), ... that all derivatives $g_r^{(\nu)}$, $\nu = 1, 2, ..., r - 1$, vanish at $V(x) = V_0(x)$, in analogy to the behavior of a polynomial

$$\sum \sum \cdots \sum k_{i,j,\cdots,l}(x_i - x_i^{(0)})(x_j - x_j^{(0)}) \cdots (x_l - x_l^{(0)}).$$

2.2. *Taylor's formula for functionals.* The functional $f\{V_1(x) + t[V(x) - V_1(x)]\}$ is for fixed $V_1(x)$ and $(V - V_1)(x)$ an ordinary function of t

$$F(t) = f\{V_1 + t(V - V_1)\}, \qquad 0 \leqslant t \leqslant 1.$$

If the functional $f\{V\}$ is differentiable up to some order $r + 1$ the derivatives of $F(t)$ up to that order exist, and we have, with θ between 0 and 1,

$$F(1) - F(0) = F'(0) + \frac{1}{2!}F''(0) + \cdots + \frac{1}{r!}F^{(r)}(0) + \frac{1}{(r+1)!}F^{(r+1)}(\theta). \tag{26}$$

The left-hand side equals $f\{V(x)\} - f\{V_1(x)\}$. The expressions $F'(0)$, $F''(0)$, ..., $F^{(r)}(0)$ are the derivatives as defined in Eqs. (16), (23), and (25). All we need is the last term in (26). Let us first consider the case $r = 1$. We set

$$V_1(x) + \theta(V - V_1)(x) = U(x), \tag{27}$$

With $(t - \theta)/(1 - \theta) = t'$ we have then

$$V_1(x) + t(V - V_1)(x) = U(x) + t'(V - U)(x),$$

and the second derivative of F with respect to t at $t = \theta$ equals the second derivative with respect to t' at $t' = 0$, divided by $(1 - \theta)^2$:

$$F''(\theta) = \frac{1}{(1-\theta)^2}\left[\frac{d^2}{dt'^2} f\{U(x) + t'(V - U)(x)\}\right]_{t'=0},$$

and since $V - U = (1 - \theta)(V - V_1)$ we obtain

$$F''(\theta) = \iint f''\{U(x), \xi_1, \xi_2\}\, dT(\xi_1)\, dT(\xi_2)$$

with $T = V - V_1$,

We state: *If the statistical function $f\{V(x)\}$ is twice differentiable on the "segment" from $V_1(x)$ to $V(x)$ the Taylor formula holds in the form*

$$\begin{aligned} f\{V(x)\} - f\{V_1(x)\} &= \int f'\{V_1(x), \xi\}\, dT(\xi) \\ &+ \tfrac{1}{2} \iint f''\{U(x), \xi_1, \xi_2\}\, dT(\xi_1)\, dT(\xi_2), \end{aligned} \tag{28}$$

where $U(x)$ is defined by (27).

For general r we have the formula

$$\begin{aligned} &f\{V(x)\} - f\{V_1(x)\} \\ &= \sum_{\rho=1}^{r} \frac{1}{\rho!} \int \cdots \int f^{(\rho)}\{V_1(x), \xi_1, ..., \xi_\rho\}\, dT(\xi_1) \cdots dT(\xi_\rho) \\ &+ \frac{1}{(r+1)!} \int \cdots \int f^{(r+1)}\{U(x), \xi_1, ..., \xi_{r+1}\}\, dT(\xi_1) \cdots dT(\xi_{r+1}). \end{aligned} \tag{29}$$

Here the ρth term ($\rho = 1, 2, ..., r$) on the right-hand side can be written

$$\frac{1}{\rho!} \int \cdots \int k(\xi_1, \xi_2, ..., \xi_\rho)\, dT(\xi_1)\, dT(\xi_2) \cdots dT(\xi_\rho), \tag{29'}$$

where $k(\xi_1, \xi_2, ..., \xi_\rho) = f^{(\rho)}\{V_1(x), \xi_1, \xi_2, ..., \xi_\rho\}$. It is seen that (with ρ for r) and except for the factor $1/\rho!$, this is the same as (17); the $k(\xi_1, ..., \xi_\rho)$ can be considered symmetric on account of Volterra's result regarding the symmetry of the "mixed" derivatives of functionals. We have found that for an expression $g_\rho\{V(x)\}$ of the form (17), i.e., a polynomial in $V - V_1$ of degree ρ all derivatives up to the order $\rho - 1$ vanish at $V = V_1$. The derivative of order ρ equals [see Eq. (18″)] $k(\xi_1, \xi_2, ..., \xi_\rho)$ (which is a "constant" since the ξ_1, ξ_2, ... play the role of "subscripts"). The complete analogy of (29) with the Taylor expansion of a function of several variables is thus recognized.

We shall be particularly interested in cases where the $(r + 1)$th derivative is bounded in a way similar to (24). Let $\psi(x)$ be a positive function. We say that $f\{V(x)\}$ *has a bounded* $(r+1)$th *derivative* at V_1 if for all $V(x)$ of the set J under consideration the inequality holds

$$\left| \iint \cdots \int f^{(r+1)}\{V_1(x), \xi_1, \xi_2, \ldots, \xi_{r+1}\}\, dT(\xi_1)\, dT(\xi_2) \cdots dT(\xi_{r+1}) \right| \leqslant \left| \int T^2(x)\psi(x)\, dx \right|^{(r+1)/2}, \tag{24'}$$

where $T = V - V_1$. This reduces to (24) for $r = 1$.

Before applying these definitions and results to our main problem, the study of the asymptotic distribution of statistical functions, we consider the laws of large numbers for statistical functions.

B. The Laws of Large Numbers (Sections 3 and 4)

3. The First Law of Large Numbers for Statistical Functions

3.1. *Historical comments.* The simplest law of large numbers is Bernoulli's theorem, 1713 (see Chapter IV). If in n simple alternatives with probability p for "success" or "event," the event has happened x times out of n, then

$$\lim_{n\to\infty} \Pr\{| \frac{x}{n} - p | > \epsilon\} = 0. \tag{30a}$$

In 1837 Poisson admitted varying probabilities p_ν, $\nu = 1, 2, \ldots$ of "success." With $\bar{p}_n = (1/n)(p_1 + \ldots + p_n)$, the arithmetical mean of the $p_1, p_2, \ldots, p_n$, we have

$$\lim_{n\to\infty} \Pr\{| \frac{x}{n} - \bar{p}_n | > \epsilon\} = 0. \tag{30b}$$

In 1846 Tchebycheff established the following generalization (see Chapter V). Let $P_\nu(x)$, $\nu = 1, 2, \ldots$ denote fairly general distributions which in n trials gave the sum $x_1 + \ldots + x_n = x$ and let $\bar{P}_n(x) = (1/n)(P_1(x) + \ldots + P_n(x))$. Then, under certain conditions concerning the behavior of $P_\nu(x)$ as $|x| \to \infty$:

$$\lim_{n\to\infty} \Pr\{\left| \frac{x}{n} - \int x\, d\bar{P}_n(x) \right| > \epsilon\} = 0. \tag{30c}$$

In 1936 v. Mises (*B*) proved the following generalization.[1] With the

[1] See partial results in R. v. Mises [21], pp. 192-197.

notation of Tchebycheff's problem let $f\{V(x)\}$ be a statistical function, continuous at the point $\bar{P}_n(x)$, and $S_n(x)$ the sample distribution of the first n results. Then

$$\lim_{n\to\infty} \Pr\{|f\{S_n(x)\} - f\{\bar{P}_n(x)\}| > \epsilon\} = 0 \tag{30d}$$

holds without any restriction if the $P_\nu(x)$ are uniformly bounded (see p. 619) while in the case of unbounded $P_\nu(x)$ a restriction involving a continuity measure $\psi(x)$ (see p. 619) must hold.[2] Equation (30d) will be proved in Sections 3.3 and 3.4.

Other related results, in particular the strong law of large numbers, and certain extensions to dependent trials, may also be generalized to statistical functions (Sections 3.5 and 3.6).

In another direction are generalizations of the "second law of large numbers," the inference law (see Chapter VII) which we shall consider in Section 4. These theorems are hardly mentioned in books on probability calculus since without the concept of a collective (or an equivalent concept) they are often not clearly distinguished from those of the first group.

3.2. *Two lemmas.* (a) The first lemma is Tchebycheff's inequality for the case of Eq. (30b). Alternatives are observed with probabilities p_ν, $\nu = 1, 2, \ldots$ for "event"; we set $\bar{p}_n = 1/n \sum_{\nu=1}^{n} p_\nu$; $x = zn$ is the number of observed events in trials. Then, for arbitrarily small η

$$\Pr\left\{\left| z - \bar{p}_n \right| > \frac{\eta}{2}\right\} \leqslant \frac{4}{n\eta^2}\, \bar{p}_n(1 - \bar{p}_n) \leqslant \frac{1}{n\eta^2}; \tag{31}$$

this, by the way, is a complete proof of (30b).

Remember also the following form of Tchebycheff's inequality. Let $P(x)$ be a probability distribution, a an arbitrary number, and $m > 1$. Then, obviously,

$$\int |x - a|^m \, dP(x) \geqslant \eta^m \int_{|x-a|>\eta} dP(x).$$

Now $\int_{|x-a|>\eta} dP(x)$ is the probability that $|x - a|$ be greater than η and we obtain

$$\Pr\{|x - a| > \eta\} \leqslant \frac{1}{\eta^m} \int |x - a|^m \, dP(x). \tag{32}$$

[2] This sketch relates to one main line of development. Some of the ramifications are mentioned at the end of Section 3.6.

(b) The second lemma is of a geometrical nature. Let $V(x)$ and $W(x)$ be two distributions, (x_1, x_2) an interval such that $V(x_2) - V(x_1) = \delta > 0$. If we know that at the two end points

$$| W(x_1) - V(x_1) | \leqslant \frac{\eta}{2}, \qquad | W(x_2) - V(x_2) | \leqslant \frac{\eta}{2} \tag{33}$$

then

$$| V(x) - W(x) | \leqslant \delta + \frac{\eta}{2}, \qquad x_1 < x \leqslant x_2 . \tag{34}$$

For $\delta = \eta/2$, (34) gives $| V(x) - W(x) | \leqslant \eta$. This is our second lemma.

3.3. *Bounded distributions.* We assume that for any ν, $P_\nu(A) = 0$, $P_\nu(B) = 1$, hence likewise $\bar{P}_n(A) = 0$, $\bar{P}_n(B) = 1$, for any n. Denote by P_η the probability that there is a point x' in the interval AB such that

$$| S_n(x') - \bar{P}_n(x') | > \eta. \tag{35}$$

We shall show that, for any given η, $\lim_{n\to\infty} P_\eta = 0$.

In the strip between the x-axis and the line $y = 1$, limited by the verticals $x = A$ and $x = B$ we draw m horizontal lines which divide $\bar{P}_n(x)$ (which, for the moment we assume to be continuous) into $m + 1$ parts of abscissas $A < a_1 < a_2 \cdots < a_m < B$. If we take $m = [2/\eta]$, where, as usual, $[a]$ means the largest integer $\leqslant a$, then the height of each strip is $\leqslant \eta/2$ hence, the variation of $\bar{P}_n(x)$ in each of the $m + 1$ intervals less than $\eta/2$. By the second lemma of Section 3.2 the inequality (35) can hold only if for one of the a_μ, $\mu = 1, 2, \ldots, m$

$$| S_n(a_\mu) - \bar{P}_n(a_\mu) | > \frac{\eta}{2}. \tag{36}$$

For any fixed a_μ, $P_\nu(a_\mu) = Pr\{x \leqslant a_\mu\} = p_\nu$ and $1 - P_\nu(a_\mu) = Pr\{x > a_\mu\} = 1 - p_\nu$ and $nS_n(a_\mu)$ is the number of those observed results which are $\leqslant a_\mu$. Hence, we may apply (31) and obtain

$$\Pr\left\{| S_n(a_\mu) - \bar{P}_n(a_\mu) | > \frac{\eta}{2}\right\} \leqslant \frac{4}{n\eta^2} \bar{P}_n(a_\mu)[1 - \bar{P}_n(a_\mu)] \leqslant \frac{1}{n\eta^2}, \tag{37}$$

and the probability that (36) holds for at least one of the $a_1, a_2, \ldots, a_m$ (these events need not be independent) cannot surpass the sum of the probabilities for the single a_μ, $\mu = 1, 2, \ldots, m$. Therefore,

$$P_\eta \leqslant \frac{m}{n\eta^2} \leqslant \frac{2}{n\eta^3}. \tag{38}$$

If one of the a_μ is a discontinuity point of $\bar{P}_n$ there are slight modifications.
From (38) we have for x in A, B

$$\Pr\{|S_n(x) - \bar{P}_n(x)| \leqslant \eta\} = 1 - P_\eta \geqslant 1 - \frac{2}{n\eta^3}. \tag{39}$$

According to the definition of continuity of a statistical function the left-hand side of (39) is at most the probability that

$$|f\{S_n(x)\} - f\{\bar{P}_n(x)\}| \leqslant \epsilon, \tag{40}$$

and our statement (30d) is proved—for uniformly bounded distributions.

I. *Let* $P_1(x)$, $P_2(x)$, ... *be an infinite sequence of uniformly bounded distributions and let* $\bar{P}_n(x) = (1/n)\sum_{\nu=1}^{n} P_\nu(x)$. *If* $S_n(x)$ *is bounded and* $f\{S_n(x)\}$ *is continuous* at $\bar{P}_n(x)$, *then*

$$\lim_{n\to\infty} \Pr\{|f\{S_n(x)\} - f\{\bar{P}_n(x)\}| > \epsilon\} = 0. \tag{30d}$$

3.4. *Unbounded distributions.* In this case we assume that there exists a positive function $\phi(x)$ decreasing with increasing $|x|$ and tending toward zero as $|x| \to \infty$, such that for some positive X

$$\left.\begin{aligned} P_\nu(x) &\leqslant \phi(x) \qquad \text{for } x \leqslant -X \\ 1 - P_\nu(x) &\leqslant \phi(x) \qquad \text{for } x \geqslant X. \end{aligned}\right\} \quad \nu = 1, 2, \dots . \tag{41}$$

Then (41) holds also for $\bar{P}_n(x)$ and

$$\bar{P}_n(x)[1 - \bar{P}_n(x)] \leqslant \phi(x) \qquad \text{for } |x| \geqslant X. \tag{41'}$$

Let $\psi(x)$ be the continuity measure (p. 619) of $f\{V(x)\}$ and assume for the moment that, for some positive constant c

$$\lim_{|x|\to\infty} \phi(x)\psi(|x| + c) \to 0. \tag{42}$$

Let $A < -X$, $B > X$ such that for given η and c

$$\frac{\eta}{2\psi(x)} \geqslant \phi(|x| - c) \qquad \text{for} \qquad x \leqslant A, \quad x \geqslant B, \tag{43}$$

which is possible on account of (42). Denote by ψ_0 the smaller of the numbers $\psi(A)$ and $\psi(B)$. If we use then $m = [2\psi_0/\eta]$ instead of the m

of Section 3.3 we can see in exactly the same way as in 3.3 that the probability of the inequality

$$| S_n(x) - \bar{P}_n(x) | > \frac{\eta}{\psi_0} \geqslant \frac{\eta}{\psi(x)}, \qquad x \text{ in } AB \tag{44}$$

tends toward zero as $n \to \infty$.

We have still to consider the intervals $x > B$ and $x < A$. We define infinitely many abscissas $b_\kappa = B + \kappa c$, where $c > 0$, $\kappa = 1, 2, \ldots$. The variation of $\bar{P}_n(x)$ in b_κ, $b_{\kappa+1}$ is by (41) and (43) less than

$$\phi(b_\kappa) \leqslant \frac{\eta}{2\psi(b_{\kappa+1})} \leqslant \frac{\eta}{2\psi(x)} \qquad \text{for} \quad x \leqslant b_{\kappa+1}\,. \tag{45}$$

We conclude then from the second lemma of Section 3.2 that (44) holds for no point in $x > B$ if the inequality

$$| S_n(b_\kappa) - \bar{P}_n(b_\kappa) | > \frac{\eta}{2\psi(b_{\kappa+1})} \tag{46}$$

holds for no b_1, b_2, ... But the probability of (46) is by (37) and (41′) less than

$$\frac{4\psi^2(b_{\kappa+1})}{n\eta^2}\, \bar{P}_n(b_\kappa\, [1 - \bar{P}_n(b_\kappa)] \leqslant \frac{4}{n\eta^2}\, \psi^2(b_{\kappa+1})\phi(b_\kappa). \tag{47}$$

Hence the probability that (44) holds anywhere in $x > B$ tends to zero as $n \to \infty$ if the sum

$$\sum_{\kappa=1}^{\infty} \psi^2(b_{\kappa+1})\phi(b_\kappa) \leqslant \frac{1}{c} \int_B^\infty \psi^2(x + c)\phi(x)\, dx \tag{48}$$

is convergent. We argue in the same way for $x < A$ and arrive, with the previous notations, at the result:

I′. *The law of large numbers* (30d) *holds for any statistical function* $f\{V(x)\}$, *continuous with respect to a continuity measure* $\psi(x)$, *if the distributions* P_ν *are restricted by* (41′), *where* $\phi(x)$ *decreases as* $|x|$ *increases, in such a way that*

$$\int \psi^2(x + c)\phi(x)\, dx \tag{49}$$

is convergent for some $c > 0$.[3]

[3] Condition (49) replaces the provisory condition (42).

A weaker condition than (49) has been established by v. Mises, (E, pp. 14 and 15). The $\int \psi^2(x + c)\phi(x)\, dx$ which was supposed to be convergent *can be replaced for any* $m = 2, 3, 4, \ldots$ *by*

$$\int [\psi^2(x + c)\phi(x)]^{m/2}\, dx. \tag{49'}$$

To prove this, the form (32) of Tchebycheff's inequality is used. To see the implication of this improvement take, for example, $\psi(x) = | x |^{\lambda}$ and $\phi(x) = | x |^{-k}$, $\lambda > 0$, $k > 0$. Then, (49) will hold if $k - 2\lambda > 1$ while (49′) will converge if $k - 2\lambda > 2/m$ where m may be arbitrarily large. In the case where $f\{V(x)\} = \int x\, dV(x)$, conditions (49) and (49′) give with $\lambda = 1 + \epsilon$, $k > 3 + \epsilon$, and $k > 2 + \epsilon$, respectively. The latter is almost as good as the classical condition of Tchebycheff and Markov established directly for the mean value (see Chapter V, Section 3).[4]

3.5. *The strong law of large numbers* (Chapter V, Section 4) can be generalized to statistical functions.[5] The proof of this important result is given in a way similar to that of the "weak" law of large numbers if Kolomorogov's inequality (Chapter III, Section 3.2) is used instead of Tchebycheff's. The lemma which we need is actually nothing but the simplest case of the law of large numbers for alternatives (see Chapter V, Section 4.2), which we state with the notation of Chapter V.

LEMMA. *Consider* $n + r$ *alternatives with event probabilities* p_1, p_2, ..., p_{n+r}. *The probability that for an integer* ν *between* n *and* $n + r$ *at least one* $| X_\nu - \nu p_\nu | \geqslant \eta$, *where* X_ν *is the number of successes out of* ν *trials, is less than or equal to* $1/n\eta^2$, *independent of* r.

The assumptions for the distributions, bounded or unbounded, are the same as in our theorems I and I′. Using the present lemma in a way similar to that in the preceding proof we obtain the results

$\mathrm{I}_{\text{strong}}$. *Under the same conditions as for* (30d), *the probability* $P(A_{n,r})$ *that*

$$| f\{S_\nu(x)\} - f\{\bar{P}_\nu(x)\} | > \epsilon$$

for any integer $\nu = n, n + 1, \ldots, n + r$, *is for every positive* ϵ *below a bound which is independent of* r *and tends toward zero as* $n \to \infty$. With C a constant

$$P(A_{n,r}) < C/n\epsilon^2. \tag{30e}$$

[4] If all P_ν are equal then, as in the classical case, weaker restrictions result.

[5] See v. MISES (B) pp. 113 and 126.

3.6. *Generalizations.* (a) In all the results reported in this section x *may designate a vector with components* x_1, x_2, ..., x_k [v. Mises (B)]. The proofs become hardly more involved and no essentially new tool is needed. Equation (30d) reads, $\mathbf{x}$ denoting a k-dimensional vector

$$\lim_{n\to\infty} \Pr\{|f\{S_n(\mathbf{x})\} - f\{\bar{P}_n\{\mathbf{x}\}|\} > \epsilon\} = 0. \tag{30f}$$

The strong law of large numbers holds likewise for vectors.

(b) The law of large numbers for statistical functions holds in certain cases of dependent collectives, as shown by Geiringer.[6]

4. The Second Law of Large Numbers for Statistical Functions

4.1. *The problem.* Such a law, known in its simplest form as Bayes' theorem, was given in Chapter VII. We discussed there its conceptual relation to the first or "direct" law of large numbers. A generalization to statistical functions is by v. Mises.[1] It is of singular importance for statistics.

The starting point is a set M of collectives of the same type, for example, various dice with different (unknown) distributions $V(x)$. Or we may think of various urns, each containing 1000 lots; the set M consists of various arithmetical distributions each with steps equal to multiples of 1/1000. There is an *a priori distribution* P_0, *which* (cf. Chapters VII and X) *gives the chance that the distribution of the urn which has been picked out, belongs to a subset of* M. We do not stipulate that P_0 be a probability— it may be a chance. In the simplest case the $V(x)$ are all alternatives (flipping a coin) and the distribution of each alternative is characterized by the value of one parameter θ; $P_0(\theta)$ may then be the *a priori* chance that for the coin picked out the (unknown) probability of "heads" be less than or equal to θ.

Suppose we pick out one of the collectives, $V(x)$, say, and make n observations on it. [Note that as in Chapters VII and X, we deal here with a less general case than in the direct problem since we use *the same* collective for the n trials rather than performing the νth trial from a νth collective.] The result of the n trials defines a sample distribution $S_n(x)$. A typical question is as follows: Given $S_n(x)$ and P_0, what is the *a posteriori* chance that the distribution picked out belongs to some subset M' of M? We have seen that the answer to such questions

[6] Hilda Geiringer, "A generalization of the law of large numbers." *Ann. Math. Statist.* 11 (1940), pp. 393-401.

[1] R. v. Mises, (B), pp. 115, and 127.

depends on $S_n(x)$ and on P_0, but that the influence of P_0 decreases gradually as n increases and that, in the limit, the *a posteriori* chance is free of P_0, due to the phenomenon of "concentration."

We wish to study this phenomenon for statistical functions.

4.2. *Lemma and theorem.* We define the set M of distributions $V(x)$; some of the $V(x)$ may be sample distributions, others arbitrary arithmetical or continuous distributions. We consider in this subsection *bounded distributions only*. There is a statistical function $f\{V(x)\}$ continuous for all $V(x)$ of M. We remember that $f\{V(x)\}$ is continuous at $V_1(x)$ if there is an η for every ϵ such that for any $V(x)$ of M for which $|V(x) - V_1(x)| \leqslant \eta$, also $|f\{V(x)\} - f\{V_1(x)\}| \leqslant \epsilon$ holds.

We shall prove the following theorem.

II. *Let $S_n(x)$ denote the sample distribution of n observations of a collective with an unknown bounded distribution $V(x)$. If f is continuous at $S_n(x)$, then the probability*

$$\Pr\{|f\{V(x)\} - f\{S_n(x)\}| > \epsilon\} \tag{50}$$

lies below a bound which for every ϵ tends to zero as $n \to \infty$, if the following restriction holds: for sufficiently large n and any fixed x the a priori chance P_0 of the inequality $|V(x) - S_n(x)| \leqslant \eta$ is greater than $c\eta^\kappa$ where η is arbitrarily small and c and κ are positive constants, independent of x.[2]

We have to realize that, in contrast to Theorem I (first law of large numbers) $S_n(x)$ is now given for a large but fixed n. We are asking for a probability of $V(x)$, given $S_n(x)$.

Again, we base our proof on a lemma which, here, is nothing but Bayes' theorem for alternatives (see Chapter VII, Section 3).

LEMMA. *In the n-fold observation of an alternative with unknown event probability p the event happened n_1 times. Put $n_1/n = r$, and assume that for the a priori distribution $P_0(r + \eta) - P_0(r - \eta) \geqslant c\eta^\kappa$, $\kappa > 0$. Then the probability of $|p - a| > \eta$ lies below a bound, which for any η tends to zero as $n \to \infty$.*

In the derivation of Theorem II we proceed as in that of Theorem I. As in Section 3.3 the segment AB is divided into $m + 1$ intervals, $m = [2/\eta]$, such that the variation of $S_n(x)$ is less than $\eta/2$ in each interval. We know then from lemma (b) of Section 3.2 that in any inner point of such an interval

$$|V(x) - S_n(x)| \leqslant \eta \tag{51}$$

[2] In Theorem II and, in particular, in (50) the "Pr" may be interpreted as a chance. The same holds for Theorem II′, p. 641.

if $S_n(x)$ and $V(x)$ differ by less than $\eta/2$ at the end points of the interval. From the present lemma we have

$$\Pr\{|p - a| > \eta\} < L(\eta, n), \tag{52}$$

where $\lim_{n\to\infty} L(\eta, n) = 0$ for any η. Therefore, the probability that (51) be violated in one or more intervals is less than

$$\frac{2}{\eta} L\left(\frac{\eta}{n}, n\right) \tag{53}$$

and indeed this goes towards zero as $n \to \infty$. Since the $m + 1$ intervals cover the segment from A to B, (53) is also an upper bound of the probability that (51) fails to hold *anywhere* between A and B except perhaps at one or more of the $m \leqslant 2/\eta$ end points of intervals. On account of the continuity of the statistical function, (53) is also an upper bound of the probability that $|f\{V(x)\} - f\{S_n(x)\}| > \epsilon$ and (50) is proved for bounded $V(x)$.

We remember the particular case of an arithmetical collective where the result becomes a statement on functions of k variables, proved in Chapter VII, Section 3.4.

If in the n-fold observations of an arithmetical collective the attributes $c_1, c_2, \ldots, c_k$ *have appeared with relative frequencies* $n_\kappa/n = r_\kappa$, $\kappa = 1, 2, \ldots, k$ *and if f is a bounded function of the (unknown) theoretical probabilities* p_κ *of the collective, then it is to be expected with a probability tending toward one as* $n \to \infty$ *that* $f(p_1, p_2, \ldots, p_k)$ *differs from the observed* $f(r_1, r_2, \ldots, r_k)$ *arbitrarily little if both f and the a priori chance* p_0 *are continuous at the point* $(r_1, r_2, \ldots, r_k)$ and $p_0 \neq 0$ *at this point.*[3]

4.3. *An example.* Before considering the case of unbounded distributions we illustrate our results by an example. Suppose that the sex rate of newly born children has been observed in n successive years and $z_1, z_2, \ldots, z_n$ are the relative frequencies of boys out of $m_1, m_2, \ldots, m_n$ observed births. Then z_ν can have any one of the $m_\nu + 1$ values 0, $1/m_\nu$, $2/m_\nu$, ..., 1. The observed $z_1, z_2, \ldots, z_n$ define an average and a dispersion:

$$a = \frac{1}{n}(z_1 + z_2 + \cdots + z_n)$$

$$s^2 = \frac{1}{n}[(z_1 - a)^2 + (z_2 - a)^2 + \cdots + (z_n - a)^2]. \tag{54}$$

[3] This last condition can be replaced by the weaker one of Theorem II, p. 637.

Denote by m the arithmetical mean of the n numbers m_ν, viz., $mn = m_1 + m_2 + \cdots + m_n$ and form the observed Lexis number

$$L = \frac{ms^2}{a(1-a)}. \tag{55}$$

Clearly L is a statistical function since

$$a = \int x\, dS_n(x), \qquad s^2 = \int (x-a)^2\, dS_n(x) = \int x^2\, dS_n(x) - a^2. \tag{54'}$$

L is bounded and continuous for bounded $V(x)$, particularly for $S_n(x)$.

On the other hand, there are n arithmetical distributions $p_\nu(x)$, $x = 0, 1, \ldots, m_\nu$, $\nu = 1, \ldots, n$, for the probability of the result x and

$$\alpha = \frac{1}{n} \sum_{\nu=1}^{n} \sum_{x=0}^{m_\nu} p_\nu(x) \frac{x}{m_\nu}$$

$$\sigma^2 = \frac{1}{n} \sum_{\nu=1}^{n} \sum_{x=0}^{m_\nu} p_\nu(x) \left(\frac{x}{m_\nu} - \alpha\right)^2. \tag{56}$$

From α and σ^2 we form the theoretical counterpart of L, namely, Λ

$$\Lambda = \frac{m\sigma^2}{\alpha(1-\alpha)}. \tag{57}$$

Let us now state our theorems in terms of this example without restating exact conditions:

Theorem I states: *If n is sufficiently large it can be expected with near certainty that the result L of observations will be arbitrarily close to the known value of Λ.*

The strong law of large numbers asserts: *If n is sufficiently large we can expect, almost with certainty, that for any value greater than a certain n, the observed L will be arbitrarily close to the given Λ.*

The typical question of statistics is, however, answered by Theorem II. We have observed a value of L; what inference can we draw on the unknown collective which is at the basis of our trials? In particular we wish to infer from L to Λ. We can give an answer only if all $p_\nu(x)$ are equal, $p_\nu(x) = p(x)$, i.e., if the circumstances in the n years can be considered as fairly stable (the so-called Poisson case): $p_\nu(x) = p(x)$, $m_\nu = m$. Then, Theorem II says: *It is almost certain, if n is sufficiently large, that the unknown Λ is arbitrarily close to the observed L.*

This conclusion, like similar inferences from an observed value to its unknown theoretical counterpart (for example, from the correlation ratio r to the corresponding ρ), are justified only by an inference theorem[4] of the type of Theorem II. A great part of theoretical statistics based on the theory of the distributions of "statistics" like X^2, ω^2, Student's t, and the correlation coefficient r, deal with the direct problem—and so does the greater part of the present chapter. However, the inverse problem, the problem of inference (see Chapters VII and X) is the problem of statistics proper.

4.4. *Unbounded distributions.* We have still to consider the case of unbounded distributions. The adaptation to this case is particularly easy since the *given* $S_n(x)$ is by definition bounded. This $S_n(x)$ has finitely many jumps only. Therefore an X exists such that $S_n(x)$ equals 0 and 1 for $x < -X$, $x > +X$, respectively. For the unknown distribution $P(x)$ we assume that it belongs to a class of distributions $V(x)$ such that $f\{V(x)\}$ is continuous at $S_n(x)$ with respect to the measure $\psi(x) = |x|^\lambda$, and that for some $k > \lambda$

$$|x|^k P(x)(1 - P(x)) \leqslant K. \tag{58}$$

Let now ϵ be given. We assume that η is such that $|f\{S_n(x)\} - f\{P(x)\}| \leqslant \epsilon$ for $|x|^\lambda |S_n(x) - P(x)| \leqslant \eta$. For this η an $X_1 > X$ can be chosen such that simultaneously

$$P(x) < \frac{\eta}{|x|^\lambda} \quad \text{for} \quad x < -X_1, \qquad 1 - P(x) < \frac{\eta}{|x|^\lambda} \quad \text{for} \quad x > X_1. \tag{59}$$

This is possible since by (58) $P(x)$ and $1 - P(x)$ go to zero like x^{-k} for large $|x|$ and $k > \lambda$. The inequalities (59) show [since $S_n(x) = 0$ for $x < -X_1$, etc.] that for $x < -X_1$, $x > X_1$:

$$|x|^\lambda |P(x) - S_n(x)| \leqslant \eta.$$

The interval $(-X_1, +X_1)$ is bounded and if we replace in this interval $|x|^\lambda$ by $X_1{}^\lambda$ and η by $\eta X_1{}^\lambda$ in the continuity definition, we have in this interval ordinary continuity and can reason as in Section 4.2. The probability that anywhere in $-\infty$, $+\infty$

$$|x|^\lambda |P(x) - S_n(x)| > \eta \tag{60}$$

is, on account of the result for $|x| > X_1$, equal to the probability that

[4] See the beginning of Chapter X.

(60) holds anywhere between $-X_1$ and $+X_1$.[5] An upper limit of this probability is (53) with η replaced by $\eta X_1{}^\lambda$. Hence, we have proved

II′. *If n observations of one and the same collective with unknown distribution $P(x)$ have given the repartition $S_n(x)$, consider $f\{V(x)\}$, a statistical function continuous at $S_n(x)$ with respect to a measure $\psi(x) = |x|^\lambda$, $\lambda > 0$. Assume that the otherwise unknown $P(x)$ satisfies* (58) *with $k > \lambda$. For the a priori chance P_0 we make the same assumption as in Theorem II. Then, the probability*

$$\Pr\{ |f\{P(x)\} - f\{S_n(x)\}| > \epsilon\} \tag{61}$$

is below a bound which for every ϵ converges to zero as $n \to \infty$.

We note that the results of this section hold again for vectors $\mathbf{x}$; hence they apply, for example, to the correlation coefficient.

C. Statistical Functions of Type One (Sections 5 and 6)

5. Convergence Toward the Normal Distribution

5.1. *The problem.* Our aim is to consider an extension of the central limit theorem (Chapter VI). The arithmetical mean of n observed values $a = (1/n) \sum_{\nu=1}^{n} x_\nu = \int x \, dS_n(x)$ is the simplest statistical function. The central limit theorem states that under weak conditions for the $P_\nu(x)$, where $P_\nu(x)$ is the distribution for the νth trial, the asymptotic distribution of a is normal with mean and variance computed from means and variances of the $P_\nu(x)$. v. Mises has proved (1936) that the same is true of a wide class of statistical functions $f\{S_n(x)\}$, which we are going to characterize.

To see the implication of this generalization, remember, for example our study of Student's $t = (\bar{x} - a)\sqrt{n-1}/s$ in Chapter IX. We assumed in Chapter IX that the sampling was from a single normal distribution and, using the fact that the numerator of t is normally distributed and s^2 has χ^2 distribution, we found its *exact* distribution, for any n; then by a passage to the limit, we saw that t was asymptotically normal. Under similar assumptions, the asymptotic normality of the correlation coefficient r follows from its exact distribution, which R. A. Fisher established under the hypothesis of (one and the same) normal parent distribution. The present approach furnishes the asymptotic distribu-

[5] In fact, outside the interval $(-X_1, +X_1)$ the inequality (60) cannot hold since $S_n(x)$ vanishes there and (59) holds for $P(x)$.

tions directly and without assuming a normal parent distribution; in fact, some general restrictions on the behavior of the $P_\nu(x)$ at infinity are all that is needed.[1] We shall see that asymptotic normality of the distributions of a comprehensive class of statistics is a general fact, largely independent of the shape of the exact distribution of the statistics under consideration as well as of the parent distributions.

Before turning to the general case we shall consider arithmetical $P_\nu(x)$. In this case, the statistical functions reduce to functions of k variables.

5.2. *A lemma.*[2] We start with a lemma which will be needed in the present particular case as well as in the more general case. Consider an infinite sequence of collectives $K_1, K_2, \ldots$ with probability distributions $P_1(x), P_2(x), \ldots$.[3] The results $x_1, x_2, \ldots, x_n$ of the first n trials give a sampling distribution $S_n(x)$. To each function f of $S_n(x)$ we associate its expectation

$$E_n[f] = \int \cdots \int f \, dP_1(x_1) \, dP_2(x_2) \cdots dP_n(x_n). \tag{62}$$

A definition of E_n equivalent to (62) is as follows. Let

$$F_n(z) = \Pr\{f\{S_n(x)\} \leqslant z\} = \int_{(f \leqslant z)} \cdots \int dP_1(x) \cdots dP_n(x) \tag{63}$$

be the distribution of $f\{S_n(x)\}$. Then,

$$E_n[f] = \int z \, dF_n(z). \tag{64}$$

Consider now the infinite sequences of functions

$$\begin{aligned} &A_1(x_1), \quad A_2(x_1, x_2), \quad A_3(x_1, x_2, x_3), \ldots \\ &B_1(x_1), \quad B_2(x_1, x_2), \quad B_3(x_1, x_2, x_3), \ldots, \end{aligned} \tag{65}$$

where x_ι has the distribution $P_\iota(x)$, $\iota = 1, 2, \ldots$. Denote by $F_n(x)$ the distributions of $A_n(x_1, x_2, \ldots, x_n)$ and by $G_n(x)$ that of $B_n(x_1, x_2, \ldots, x_n)$

$$F_n(x) = \Pr\{A_n \leqslant x\}, \qquad G_n(x) = \Pr\{B_n \leqslant x\}. \tag{66}$$

---

[1] Some results of this general type are in Cramér [4], pp. 363-370. (Cramér quotes v. Mises' work.) See also W. G. Madow "Limiting distributions of quadratic and bilinear forms." *Ann. Math. Statist.* 11 (1940), p. 125. See also comments at the end of Section 6.

[2] See v. Mises (C); Cramér [4], p. 254.

[3] The dimensions of the collectives could be higher than one.

We assume that $\lim_{n\to\infty} G_n(x) = G(x)$, *and that for any* $l > 0$:

$$\lim_{n\to\infty} \Pr\{| A_n - B_n | \geqslant l\} = 0. \tag{67}$$

We shall prove that *at all points where* $G(x)$ *is continuous*, $F_n(x)$ *tends* toward $G(x)$.

From (66) it follows that

$$\begin{aligned} F_n(x) - G_n(x) = {} & [\Pr\{A_n \leqslant x, B_n > x\} + \Pr\{A_n \leqslant x, B_n \leqslant x\}] \\ & - [\Pr\{A_n \leqslant x, B_n \leqslant x\} + \Pr\{A_n > x, B_n \leqslant x\}]. \end{aligned}$$

Therefore

$$F_n(x) - G_n(x) = \Pr\{A_n \leqslant x, B_n > x\} - \Pr\{A_n > x, B_n \leqslant x\}. \tag{68}$$

Now, for $l > 0$,

$$\Pr\{A_n \leqslant x, B_n > x\} \leqslant \Pr\{A_n - B_n \leqslant -l\} + \Pr\{x < B_n \leqslant x + l\}.$$

In fact, the region described to the left of this inequality is contained in the sum of the two regions to the right. The last term to the right is equal to $G_n(x + l) - G_n(x)$; since $G_n(x) \to G(x)$, that difference may be replaced by $G(x + l) - G(x) + \epsilon$, where $\epsilon \to 0$ as $n \to \infty$, and since x is a continuity point of $G(x)$, the difference, which we call η, tends to zero with l. Hence

$$\Pr\{A_n \leqslant x, B_n > x\} \leqslant \Pr\{| A_n - B_n | \geqslant l\} + \eta + \epsilon, \tag{69}$$

where η can be made arbitrarily small by taking l small enough; and $\epsilon \to 0$ as $n \to \infty$. The same conclusion may be drawn for the second term $\Pr\{A_n > x, B_n \leqslant x\}$, in (68) and since probabilities are non-negative it holds for the difference and we obtain

$$| F_n(x) - G_n(x) | \leqslant \Pr\{ | A_n - B_n | \geqslant l\} + \eta + \epsilon_1 . \tag{70}$$

Now let $n \to \infty$ and then $l \to 0$, and the statement is proved.

It is convenient for our purposes to assume that, with definition (62):

$$\lim_{n\to\infty} E_n[| A_n - B_n |] = 0. \tag{71}$$

This condition implies (67). We therefore state our lemma as follows:

Let A_n , B_n *be defined by* (65) *with independent* x_1 , x_2 , ... *and* $P_\nu(x)$ be the distribution of x_ν . *Let* $F_n(x)$, $G_n(x)$ *be the distributions of* A_n *and* B_n

as in (66), *and* $E_n = \int | A_n - B_n | dV_n$, *where* $dV_n = dP_1(x) \cdots dP_n(x)$. *If then*

$$\lim_{n\to\infty} E_n = 0, \qquad \lim_{n\to\infty} G_n(x) = G(x) \tag{72}$$

then at all continuity points of $G(x)$:

$$\lim_{n\to\infty} F_n(x) = G(x). \tag{73}$$

We note that using Schwarz' inequality we can replace the condition (71) by

$$\lim_{n\to\infty} E_n[(A_n - B_n)^2] = 0. \tag{74}$$

5.3. *Some facts on alternatives.* The $P_\nu(x)$ are arithmetical mutually independent distributions with attributes $c_1, c_2, \ldots, c_k$; we denote by $P_{\nu\kappa}$ the step of $dP_\nu(x)$ at c_κ and put

$$\bar{p}_\kappa = \frac{1}{n}(p_{1\kappa} + p_{2\kappa} + \cdots + p_{n\kappa}). \tag{75}$$

In n trials, the result c_κ has been observed n_κ times; we put $n_\kappa/n = r_\kappa$, $\kappa = 1, \ldots, k$, $\sum_{\kappa=1}^{k} r_\kappa = 1$. Denote by $P_n(r_1, r_2, \ldots, r_k) \equiv P_n(r)$ the probability of the results $r_1, r_2, \ldots, r_k$ in n trials. We want the moments of first and second order of $P(r)$.[4]

$$E_n[n_\kappa] = E[nr_\kappa] = p_{1\kappa} + p_{2\kappa} + \cdots + p_{n\kappa}. \tag{76}$$

and

$$E_n[r_\kappa] = \frac{1}{n}\sum_{\nu=1}^{n} p_{\nu\kappa} = \bar{p}_\kappa, \tag{76'}$$

$$\mathrm{Var}_n[r_\kappa] = E_n[(r_\kappa - \bar{p}_\kappa)^2] = \frac{1}{n^2}\sum_{\nu=1}^{n} p_{\nu\kappa}(1 - p_{\nu\kappa}). \tag{77}$$

One may easily verify the identity

$$\frac{1}{n^2}\sum_{\nu=1}^{n} p_{\nu\kappa}(1 - p_{\nu\kappa}) = \frac{1}{n}\bar{p}_\kappa(1 - \bar{p}_\kappa) - \frac{1}{n^2}\sum_{\nu=1}^{n}(p_{\nu\kappa} - \bar{p}_\kappa)^2, \tag{78}$$

which gives the result [used already in the derivation of (31)]

$$E_n[(r_\kappa - \bar{p}_\kappa)^2] = \mathrm{Var}_n[r_\kappa] \leqslant \frac{1}{n}\bar{p}_\kappa(1 - \bar{p}_\kappa) \leqslant \frac{1}{n}\bar{p}_\kappa. \tag{79}$$

[4] The notation differs from that in Chapter XI. There we had p_κ for the present r_κ and π_κ for the present $\bar{p}_\kappa$. It would be impractical to denote here all probabilities by Greek letters.

Denote by $\mathbf{r}$ the vector with components $r_1, r_2, \ldots, r_k$, by $\mathbf{p}$ the vector with components $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_k$; then

$$(\mathbf{r} - \mathbf{p})^2 = \sum_{\kappa=1}^{k} (r_\kappa - \bar{p}_\kappa)^2$$

and from this and (79), and $\Sigma \bar{p}_\kappa = 1$:

$$E_n[(\mathbf{r} - \mathbf{p})^2] \leqslant \frac{1}{n}. \tag{80}$$

5.4. *Limit theorem for arithmetic distributions.*[5] Consider a statistical function $f\{S_n(x)\}$. It will be a function of the relative frequencies $r_1, r_2, \ldots, r_k$. The region D of the $r_\kappa = n_\kappa/n$ is

$$D: \quad r_1 \geqslant 0, \quad r_2 \geqslant 0, \ldots, r_k \geqslant 0, \quad r_1 + r_2 + \cdots + r_k = 1. \tag{81}$$

The same D contains the $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_k$ for any n. From a certain n on the r_κ and the $\bar{p}_\kappa$ will be in a subregion D_1 of D which is bounded away from the boundary of D.[6]

With $f\{S_n(x)\} = f(r_1, r_2, \ldots, r_k)$, we denote by $f_\kappa = \left(\frac{\partial f}{\partial r_k}\right)_{\mathbf{r}=\mathbf{p}}$, the derivative of f with respect to r_κ at the point $\mathbf{r} = \mathbf{p}$ [where $\mathbf{p} = (\bar{p}_1, \ldots, \bar{p}_k)$ depends of course on n]. We assume that $f(r_1, r_2, \ldots, r_k) \equiv f(r)$ is bounded for the r_κ in D and that it has continuous and bounded derivatives of first and second order at least, from a certain n on, in the subspace D_1. We set up Taylor's formula of second order (as in Chapter XI, Section 8) and have, writing now r and $\bar{p}$ for $r_1, r_2, \ldots, r_k$ and $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_k$

$$f(r) - f(\bar{p}) = \sum_{\kappa=1}^{k} (r_\kappa - \bar{p}_\kappa) f_\kappa + R, \tag{82}$$

where

$$R = \tfrac{1}{2} \sum_{\kappa,\lambda}^{1\cdots k} (r_\kappa - \bar{p}_\kappa)(r_\lambda - \bar{p}_\lambda) f_{\kappa\lambda}(r'), \tag{83}$$

r' being some point in D_1-space between r and $\bar{p}$ and the $f_{\kappa\lambda}(r')$ are bounded by hypothesis.

Let σ_n be a constant to be defined presently and put

$$A_n = \frac{n}{\sigma_n}[f(r) - f(\bar{p})], \qquad B_n = \frac{n}{\sigma_n} \sum_{\kappa=1}^{k} (r_\kappa - \bar{p}_\kappa) f_\kappa; \tag{84}$$

[5] In this proof no functionals are needed.

[6] That means we avoid a situation of "rare events."

then from (82)

$$A_n - B_n = \frac{n}{\sigma_n} R. \tag{85}$$

We wish to show

(1) That the distribution $G_n(x)$ of B_n tends toward the Gaussian $\Phi(x)$

(2) That $\lim_{n\to\infty} E[|A_n - B_n|] = 0$.

It will then follow from the lemma of Section 5.2 that

$$F_n(x) = \Pr\left\{\frac{n}{\sigma_n}(f(r) - f(\bar{p})) \leqslant x\right\} \tag{86}$$

tends toward $\Phi(x)$.

To prove (1) we construct a sequence of arithmetical distributions with jumps of magnitudes $p_{\nu\kappa}$ and the abscissas (attributes) f_κ, $\kappa = 1, 2, \ldots, k$, $\nu = 1, 2, \ldots$. Let α_ν and ρ_ν^2 be mean value and variance of the νth distribution

$$\alpha_\nu = \sum_{\kappa=1}^{k} f_\kappa p_{\nu\kappa}, \qquad \rho_\nu = \sum_{\kappa=1}^{k} (f_\kappa - \alpha_\nu)^2 p_{\nu\kappa} = \sum_{\kappa=1}^{k} f_\kappa^2 p_{\nu\kappa} - \alpha_\nu^2. \tag{87}$$

The average x/n of the results in n trials is

$$\frac{x}{n} = \sum_{\kappa=1}^{k} f_\kappa r_\kappa. \tag{88}$$

We denote by β_n and σ_n^2/n^2 the mean value and variance of x/n. They are

$$\beta_n = \frac{1}{n}\sum_{\nu=1}^{n} \alpha_\nu = \frac{1}{n}\sum_{\nu=1}^{n}\sum_{\kappa=1}^{k} f_\kappa p_{\nu\kappa} = \sum_{\kappa=1}^{k} f_\kappa \bar{p}_\kappa,$$
$$\frac{\sigma_n^2}{n^2} = \frac{1}{n^2}\sum_{\nu=1}^{n} \rho_\nu^2 = \frac{1}{n^2}\sum_{\nu=1}^{n}\left[\sum_{\kappa=1}^{k} f_\kappa^2 p_{\nu\kappa} - \left(\sum_{\kappa=1}^{k} f_\kappa p_{\nu\kappa}\right)^2\right]. \tag{89}$$

We know that a sum of independent chance variables tends under certain general conditions toward the Gaussian Φ (Chapter VI). Simple sufficient conditions are, for example, the boundedness of the absolute moments of order three and the existence of a lower bound for the variances. Since we assumed the f_κ to be bounded, $f_\kappa < M$, we have $\sum_{\kappa=1}^{k} |f_\kappa|^3 p_\kappa < M^3$. Let us assume, in addition, that two subscripts α, β exist, such that from a certain n on

$$p_{n\alpha} > \eta, \qquad p_{n\beta} > \eta, \qquad |f_\alpha - f_\beta| > \eta. \tag{90}$$

Then we have

$$\rho_\nu^2 \geqq p_{\nu\alpha}(f_\alpha - \alpha_\nu)^2 + p_{\nu\beta}(f_\beta - \alpha_\nu)^2 \geqslant \eta[(f_\alpha - \alpha_\nu)^2 + (f_\beta - \alpha_\nu)^2]$$
$$\geqslant \left[\left(\frac{f_\alpha - f_\beta}{2}\right)^2 + \left(\frac{f_\alpha + f_\beta}{2} - \alpha_\nu\right)^2\right] 2\eta > \frac{\eta^3}{2}. \tag{91}$$

Hence, both sufficient conditions are satisfied. It follows that the x/n of (88) is asymptotically normal with mean value β_n and variance σ_n^2/n^2 as in (89). Now the B_n of (84) is by (88) and (89)

$$B_n = \frac{n}{\sigma_n}\left[\frac{x}{n} - \sum_{\kappa=1}^{k} f_\kappa \bar{p}_\kappa\right] = \frac{n}{\sigma_n}\left[\frac{x}{n} - \beta_n\right]. \tag{92}$$

Therefore the limit distribution of B_n equals $\Phi(x)$.

We have still to show that the expectation of $| A_n - B_n |$ tends to zero as $n \to \infty$. This has actually been shown in Chapter XI, p. 610 ff. We sketch the proof. From (83)

$$2R = \sum_{\kappa,\lambda}^{1\cdots k} (r_\kappa - \bar{p}_\kappa)(r_\lambda - \bar{p}_\lambda) f_{\kappa\lambda}(r'),$$

where r' must be in D_1. Since all second derivatives are bounded, $| f_{\kappa\lambda} | < M_2$, we obtain, using $E[| ab |] \leqslant \frac{1}{2}E[a^2] + \frac{1}{2}E[b^2]$

$$E_n[| R |] \leqslant \tfrac{1}{2} M_2 E_n \left| \sum_{\kappa,\lambda}^{1\cdots k} | r_\kappa - \bar{p}_\kappa | \cdot | r_\lambda - \bar{p}_\lambda | \right| \leqslant \frac{k}{2} M_2 E_n \left| \sum_{\kappa=1}^{k} (r_\kappa - \bar{p}_\kappa)^2 \right|. \tag{93}$$

With the previous abbreviations: $\sum_{\kappa=1}^{k} (r_\kappa - \bar{p}_\kappa)^2 = (\mathbf{r} - \mathbf{p})^2$ and using (80) we obtain

$$E_n[| A_n - B_n |] \leqslant \frac{n}{\sigma_n} \frac{k}{2} M_2 E_n[(\mathbf{r} - \mathbf{p})^2] \leqslant \frac{nk}{\sigma_n} \frac{M_2}{2n} = \frac{kM_2}{2\sigma_n}. \tag{93'}$$

Now from (91) $\sigma_n^2 > (n/2)\,\eta^3$ or $1/\sigma_n < \sqrt{2/n\eta^3}$, and it follows that

$$\lim_{n\to\infty} E_n[| A_n - B_n |] = 0.$$

We have thus proved the theorem.

THEOREM A. *Let the statistical function $f(r)$ be a function of the relative frequencies r_κ, $\kappa = 1, \ldots, k$ such that* (1) *$f(r)$ is bounded in the region D defined in* (81), (2) *$f(r)$ has continuous and bounded derivatives* *of first and second order in the subregion D_1 of D,* (3) *there are two subscripts α, β, and a positive η such that* (90) *holds.*

Then with the σ_n of (89) *and the notation of* (82)

$$\lim_{n\to\infty} \Pr\left(\frac{n}{\sigma_n}[f(r) - f(\bar{p})] \leqslant x\right) = \Phi(x). \tag{94}$$

The present theorem is contained in the more general theorem of the next section. We proved it separately, however, since the non-elementary concepts of the differentiability of statistical functions were not needed.

We add the following remark. The results contained in Eqs. (41) and (42), Chapter XI and part of the present Theorem A are comparable. In the former case, we computed the limit of the variance of $f(r)$ under the assumption that all n trials were subject to the same distribution. Here, we have computed the asymptotic distribution of $f(r)$—and its variance under conditions (1) to (3), where, in particular, (3) was not needed in Chapter XI. On the other hand, the proof of Chapter XI required conditions for the third derivatives $f_{\kappa\lambda\rho}$. (See Chapter XI, and Section 8.1).

6. Convergence toward the Gaussian Distribution. General Case

6.1. *Preparations.* We abandon now the restriction to arithmetical distributions with k attributes. In line with the lemma of Section 5.2 we shall define the analog of the expressions A_n and B_n and show that the conditions (1) and (2) of p. 646 are satisfied.

First we transfer the simple results of Section 5.3 to continuous distributions. Let $P_\nu(x)$ be the probability of a result $\leqslant x$ in the νth observation, and $nS_n(x)$, the number of results $\leqslant x$ in n trials. Let x be a fixed value; replace all attributes less than or equal to x by 1, the others by 0. Then $P_\nu(x)$ and $1 - P_\nu(x)$ are probabilities of the results 1 and 0, respectively, and $nS_n(x)$ is the sum of the results of n trials. We may then apply (76′) and (77) and obtain

$$E_n[S_n(x)] = \frac{1}{n}[P_1(x) + \cdots + P_n(x)] = \bar{P}_n(x)$$

$$\operatorname{Var}_n[S_n(x)] = \frac{1}{n^2}\sum_{\nu=1}^{n} P_\nu(x)(1 - P_\nu(x)), \tag{77′}$$

and (79) becomes

$$E_n[S_n(x) - \bar{P}_n(x)^2] \leqslant \frac{1}{n}\bar{P}_n(x)[1 - \bar{P}_n(x)]. \tag{79′}$$

It will be practical to use occasionally the notation

$$\frac{1}{n}\sum_{\nu=1}^{n} P_\nu(x)P_\nu(y) = \bar{\bar{P}}_n(x, y)$$

which complements the notation $(1/n)\sum_{\nu=1}^{n} P_\nu(x) = \bar{P}_n(x)$.

Another expectation will also be needed. Let $\psi(x)$ be a non-negative function of x and consider the integral

$$I = \int \psi(x)[S_n(x) - \bar{P}_n(x)]^2\, dx. \tag{95}$$

We know from Chapter IX that with

$$\omega^2 = \int \psi(x)[S_n(x) - P(x)]^2\, dx$$

we have

$$E[\omega^2] = \frac{1}{n}\int \psi(x)P(x)(1 - P(x))\, dx.$$

If the $P_\nu(x)$ change from trial to trial, the $P(x)[1 - P(x)]$ in the preceding equation is replaced by $(1/n)\sum_{\nu=1}^{n} P_\nu(x)[1 - P_\nu(x)]$ and this is less than $\bar{P}_n(x)[1 - \bar{P}_n(x)]$. Hence,

$$\begin{aligned} E_n[I] &= E_n\Big[\int \psi(x)(S_n(x) - \bar{P}_n(x))^2\, dx\Big] \qquad (96)\\ &\leqslant \frac{1}{n}\int \psi(x)\bar{P}_n(x)[1 - \bar{P}_n(x)]\, dx. \end{aligned}$$

We assume that this expression exists.

Consider the Taylor expansion for statistical functions, Eq. (28). We write $S_n(x)$ for $V(x)$, $\bar{P}_n(x)$ for $V_1(x)$, and $T_n(x) = S_n(x) - \bar{P}_n(x)$ for $T(x)$. We also replace $U(x)$, defined in Eq. (27), by $P(x)$, a "point" on the segment from $S_n(x)$ to $\bar{P}_n(x)$. Multiplying both sides by n/σ_n, where σ_n will be given explicitly in Eq. (103), we have

$$\begin{aligned} \frac{n}{\sigma_n}[f\{S_n(x)\} - f\{\bar{P}_n(x)\}] &= \frac{n}{\sigma_n}\int f'\{\bar{P}_n(x), \xi\}\, dT_n(\xi)\\ &+ \frac{1}{2}\frac{n}{\sigma_n}\iint f''\{P(x), \xi_1, \xi_2\}\, dT_n(\xi_1)\, dT_n(\xi_2). \qquad (97) \end{aligned}$$

We set

$$A_n = \frac{n}{\sigma_n}[f\{S_n(x)\} - f\{\bar{P}_n(x)\}], \qquad B_n = \frac{n}{\sigma_n}\int f'\{\bar{P}_n(x), \xi\}\, dT_n(\xi) \tag{98}$$

$$A_n - B_n = \frac{n}{\sigma_n}\cdot\frac{1}{2}\iint f''\{P(x), \xi_1, \xi_2\}\, dT_n(\xi_1)\, dT_n(\xi_2). \tag{99}$$

Let us remember the Liapounoff condition (Chapter VI, Section 7): $\sum_{\nu=1}^{n} M_{\nu}^{|k|}/s_n^{\,k} \to 0$, where $k > 2$, $M_{\nu}^{|k|}$ the kth absolute central moment of $P_{\nu}(x)$ and $s_n^{\,2}$ its variance. This condition is satisfied if *we assume that for all $P_{\nu}(x)$ the absolute moments of order $(2 + \epsilon)$ exist for some $\epsilon > 0$ and are bounded and that $\sum_{\nu=1}^{n} r_{\nu}^{2} = s_n^{\,2}$ increases more strongly than $n^{2/(2+\epsilon)}$.*

We remember that the conditions for the central limit theorem simplify if all P_{ν} are equal. Then it is sufficient to assume that the common distribution $P_{\nu} = P$ has a finite moment ρ^2 of second order. This follows from the fact that (L), the Lindeberg condition $\lim_{n\to\infty} (1/s_n^{\,2}) \sum_{\nu=1}^{n} \int_{|x|>\epsilon s_n} x^2 \, dP_{\nu}(x) = 0$ is satisfied in this case. [Likewise, (L) is satisfied if all P_{ν} are bounded.]

6.2. *Proof of the main theorem.* We prove now that the distribution of B_n approaches a normal distribution. For a fixed distribution $\bar{P}_n(x)$ the expression $f'\{\bar{P}_n(x), y\}$ is a function of y and we have

$$B_n \frac{\sigma_n}{n} = \frac{1}{n} \sum_{\nu=1}^{n} f'\{\bar{P}_n(x), x_{\nu}\} - \frac{1}{n} \sum_{\nu=1}^{n} \int f'\{\bar{P}_n(x), \xi\} \, dP_{\nu}(\xi). \tag{100}$$

Now we apply the same idea as on p. 646: we consider the $f'\{\bar{P}_n(x), x_{\nu}\}$ as new attributes (results) in place of the results x_{ν}. The first sum to the right in (100) is then the sum of n independent random variables and each term of the second sum in (100) is the mean value of one of these variables. Under appropriate conditions, this sum will be asymptotically normal. We use the abbreviations

$$f_{\nu} = f'\{\bar{P}_n(x), x_{\nu}\}, \qquad \alpha_{\nu} = \int f'\{\bar{P}_n(x), \xi\} \, dP_{\nu}(\xi) \tag{101}$$

and have simply

$$B_n = \frac{1}{\sigma_n} \sum_{\nu=1}^{n} (f_{\nu} - \alpha_{\nu}), \tag{102}$$

where the right-hand side is a sum of independent random variables with zero expectations. The set of sufficient conditions of Section 6.1 now reads as follows. Put, with α_{ν} of (101),

$$\rho_{\nu}^{2} = \int [f'\{\bar{P}_n(x), \xi\} - \alpha_{\nu}]^2 \, dP_{\nu}(x), \qquad \sigma_n^{2} = \rho_1^{2} + \rho_2^{2} + \cdots + \rho_n^{2} \tag{103}$$

$$M_{\nu} = \int |f'\{\bar{P}_n(x), \xi\} - \alpha_{\nu}|^{2+\epsilon} \, dP_{\nu}(\xi). \tag{104}$$

We have seen that for the validity of Liapounoff's condition it is sufficient that for some $\epsilon > 0$, the M_ν are bounded and that $\lim_{n\to\infty} \sigma_n^2/n^{2/(2+\epsilon)} \to \infty$. Under these conditions the distributions $G_n(x)$ of B_n [defined by (98) or (102) with the σ_n of (103)] converges toward $\Phi(x)$. Note that all points of $G(x) = \Phi(x)$ are continuity points.

It remains to estimate $E[|A_n - B_n|]$. We remember condition (24), (p. 627) regarding the boundedness of a second derivative. We shall now assume that at any point $P(x)$ of a convex neighborhood of $\bar{P}_n(x)$, $f''\{P(x), \xi_1, \xi_2\}$ is bounded in this sense: There exists a non-negative $\psi(x)$ such that with $T_n(x) = S_n(x) - \bar{P}_n(x)$

$$\left| \iint f''\{P(x), \xi_1, \xi_2\} \, dT_n(\xi_1) \, dT_n(\xi_2) \right| \leqslant \int \psi(x) T_n^2(x) \, dx \tag{105}$$

or, by (99),

$$| A_n - B_n | \leqslant \frac{n}{2\sigma_n} \int \psi(x) T_n^2(x) \, dx. \tag{106}$$

The right-hand side or (105) is, except for the factor $n/2\sigma_n$, the I of (95), for which we have seen in *Eq.* (96) that

$$E_n[I] \leqslant \frac{1}{n} \int \psi(x) \bar{P}_n(x)[1 - \bar{P}_n(x)] \, dx. \tag{96}$$

Therefore,

$$E_n[| A_n - B_n |] \leqslant \frac{1}{2\sigma_n} \int \psi(x) \bar{P}_n(x)(1 - \bar{P}_n(x)) \, dx. \tag{107}$$

Hence, if the right-hand side in (107) tends to zero, $E_n[|A_n - B_n|]$ will tend to zero. We have thus reached the following generalization of the central limit theorem [v. Mises (C)][1].

THEOREM B. *The asymptotic distribution of the variate*

$$\frac{n}{\sigma_n} [f\{S_n(x)\} - f\{\bar{P}_n(x)\}] \tag{108}$$

with σ_n^2 *defined in* (103) *equals* $\Phi(x)$, *if*

(a) *The statistical function* $f\{V(x)\}$ *is twice differentiable in a convex neighborhood* J *of* $\bar{P}_n$. This J must contain all possible $S_n(x)$ and the $\bar{P}_n(x)$, at least from some n on.

[1] The result has been generalized to distributions $V(\mathbf{x})$ by Y. GARTI, *Rev. Math. Union Interbalcanique* **3** (1940), pp. 1-19.

(b) *The first derivative* $f'\{\bar{P}_n(x), \xi\}$ *is such that for the random variable* $f_\nu - \alpha_\nu$, *defined in* (101) *with moments* (103) *and* (104) *sufficient conditions for the central limit theorem are satisfied.*

(*c*) *The second derivative* $f''\{\bar{P}_n(x), \xi_1, \xi_2\}$ *satisfies the condition* (105) *with a non-negative* ψ *for which the right-hand side in* (107) *tends to zero.*

If all $P_\nu(x)$ are equal to the same $P(x)$ condition (b) simplifies. It is then sufficient that, with $\alpha = \int f'\{P(x), \xi\}\, dP(x)$ the variance $\rho^2 = \int [f'\{P(x), \xi\} - \alpha]^2\, dP(x)$ exists.

6.3. *Examples*. (a) The linear functions $\int a(x)\, dV(x)$ satisfy our conditions if the conditions of the central limit theorem are satisfied. In fact, the first derivative is $a(x)$, independent of $V(x)$, the second is zero. Take, for example, $a(x) = (x - d)^\mu$, d constant. We set

$$c_\mu = \int (x - d)^\mu\, dS_n(x) = \frac{1}{n}[(x_1 - d)^\mu + \cdots + (x_n - d)^\mu],$$

and assume that all P_ν are the same.[2] We compute ρ^2 from (103) writing γ_μ for the $\alpha_\nu = \alpha$ in (101)

$$f' = (x - d)^\mu, \qquad \int (x - d)^\mu\, dP = \gamma_\mu, \qquad \rho^2 = \int f'^2\, dP - \gamma_\mu{}^2 = \gamma_{2\mu} - \gamma_\mu{}^2.$$

Therefore,

$$\sqrt{n}\,\frac{c_\mu - \gamma_\mu}{\sqrt{\gamma_{2\mu} - \gamma_\mu{}^2}} \tag{110}$$

is asymptotically normal with parameters 0 and 1.

(b) Consider a non-linear function of the type (20)

$$f = F(A, B, C, \ldots), \qquad A = \int a(x)\, dV(x), \qquad B = \int b(x)\, dV(x), \ldots. \tag{111}$$

Assume that F is continuous and has bounded first and second derivatives with respect to $A, B, \ldots$ at least in the neighborhood of $V(x) = \bar{P}_n(x)$ from a certain n on. The central moments belong to this type since $m_r = \int (x - a)^r\, dS_n(x)$, with $a = \int x\, dS_n(x)$, is a function of zero moments of orders $\nu = 1, 2, \ldots, r$. Therefore, functions of these central moments (Lexis quotient, Student's t, correlation coefficient, etc.) belong to the above type.

[2] If the P_ν are not assumed to be the same, the random variables $a(x)$ with distributions $P_\nu(x)$ must satisfy sufficient conditions for the validity of the central limit theorem.

From (111) we have (writing y, z, ... for ξ_1, ξ_2, ...) [see (20″), p. 625)

$$f' = \frac{\partial F}{\partial A} a(y) + \frac{\partial F}{\partial B} b(y) + \cdots . \tag{111'}$$

Our second hypothesis (b) of the theorem will be satisfied if all $a(x)$, $b(x)$, ... satisfy the conditions of the central limit theorem. We consider now condition (c) of the theorem. Here f'' is a sum of terms of the form

$$\frac{\partial^2 F}{\partial A^2} a(y)a(z), \qquad \frac{\partial^2 F}{\partial A \partial B} a(y)b(z), \ \ldots .$$

and the $\partial^2 F/\partial A^2$, $\partial^2 F/\partial A \, \partial B$, ... are bounded. Consider, according to (105)

$$\iint a(y)b(z) \, dT_n(y) \, dT_n(z) = \int a(x) \, dT_n(x) \int b(x) \, dT_n(x).$$

Suppose that the derivatives $a'(x)$, $b'(x)$, ... possess a majorant $\psi_1(x)$

$$| \, a'(x) \, | < \psi_1(x), \qquad | \, b'(x) \, | < \psi_1(x), \ \ldots . \tag{112}$$

Integration by parts gives

$$\int a(x) \, dT_n(x) = - \int a'(x) T_n(x) \, dx$$

and by Schwarz' inequality with $\psi(x) > 0$

$$\left| \int a'(x) T_n(x) \, dx \right|^2 \leqslant \int \frac{a'(x)^2}{\psi(x)} \, dx \int \psi(x) T_n^2(x) \, dx \tag{113}$$

and, therefore, using (112)

$$\left| \iint a(y)b(z) \, dT_n(y) \, dT_n(z) \right| \leqslant \int \frac{\psi_1^2(x)}{\psi(x)} \, dx \cdot \int \psi(x) T_n^2(x) \, dx.$$

Therefore our hypothesis (c) will hold if

$$\int \frac{\psi_1^2}{\psi} \, dx \tag{114}$$

converges, ψ being a function for which the right-hand side of (107) goes to zero, or, for which the integral in (107) is bounded. In this form the condition (c) is easily checked.

Let us compute the α_ν and ρ_ν^2. Denote by F_1, F_2, ... the $\partial F/\partial A$,

$\partial F/\partial B$, ... with the A, B, ... replaced after differentiation by their respective values at $\bar{P}_n(x)$, which we call for brevity $\tilde{A}$, $\tilde{B}$, ... Considering A and B only we obtain

$$\alpha_\nu = \int [F_1 a(x) + F_2 b(x)]\, dP_\nu$$

$$\begin{aligned}\rho_\nu^2 &= \int [F_1 a(x) + F_2 b(x)]^2\, dP_\nu - \alpha_\nu^2 \\ &= F_1^2\Big[\int a(x)^2\, dP_\nu - \tilde{A}^2\Big] + 2F_1F_2\Big[\int a(x)b(x)\, dP_\nu - \tilde{A}\tilde{B}\Big] \\ &\quad + F_2^2\Big[\int b(x)^2\, dP_\nu - \tilde{B}^2\Big]\,. \end{aligned} \tag{115}$$

We investigate condition (c) of the theorem, assuming that $f\{S_n(x)\}$ is a function of several moments—for example of a and s^2, like the Lexis quotient or like Student's t; hence $a(x) = x$, $b(x) = (x - a)^2$. We can then take $\psi_1(x) = c + 2|\,x\,|$, $c > 0$. The integral (114) will converge if ψ increases at infinity more strongly than $|\,x\,|^{3+\epsilon}$, $\epsilon > 0$. Consider (107); assume that $|\,x\,|^{4+\epsilon} P_\nu(x)$ for $x \to -\infty$, and similarly $|\,x\,|^{4+\epsilon}(1 - P_\nu(x))$ for $x \to \infty$, remain below a bound independent of ν. Then, the integrals (107) converge uniformly with respect to n and condition (c) is satisfied.

If f depends on moments up to the order m we see in the same way that it is sufficient that $P_\nu(x)\,|\,x\,|^{2m+\epsilon}$ remain bounded with $\epsilon > 0$, as $x \to -\infty$ and similarly for $1 - P_\nu(x)$.

(c) As a next example we could establish the asymptotic distribution of the maximum-likelihood estimate which we obtained in Chapter X, Section 6.4 for the case of arithmetical distributions.

(d) Suppose that f depends on the average a and certain central moments m_ρ about a:

$$f\{S_n(x)\} = F(a, m_2, m_3, \ldots, m_r) \tag{116}$$

We assume $P_\nu(x) = P(x)$ and with the notations and computations as above we obtain

$$\begin{aligned}\gamma^2 = \rho^2 &= F_1^2\mu_2 + 2F_1 \sum_{i=1}^{r} F_i(\mu_{i+1} - i\mu_2\mu_{i-1}) \\ &\quad + \sum_{i=2}^{r} F_i^2(\mu_{2i} - \mu_i^2 - 2i\mu_{i-1}\mu_{i+1} + i^2\mu_2\mu_{i-1}^2) \\ &\quad + 2 \sum_{i<j}^{2\cdots r} F_iF_j(\mu_{i+j} - \mu_i\mu_j - i\mu_{i-1}\mu_{j+1} - j\mu_{j-1}\mu_{i+1} + ij\mu_2\mu_{i-1}\mu_{j-1}). \end{aligned} \tag{117}$$

For the γ^2 on the left-hand side we may write n Var$[f]$ since

$$\mathrm{Var}[f] = \sigma_n^2/n^2 = n\gamma^2/n^2 = \gamma^2/n. \tag{118}$$

The result is that the statistical function (116) is asymptotically normal with mean $f(\alpha, \mu_2, \mu_3, \ldots, \mu_r)$ and variance (118) if the right-hand side in (117) exists and if (with m replaced by r) the condition immediately preceding (c) holds.

Other statistical functions which are asymptotically normal are the median and other quantiles.

6.4. *Comments.* We have thus found that the normal distribution is indeed the "normal" or general asymptotic distribution ("general" in the same sense as when we say a function $F(x)$ has "in general" a non-zero derivative at a point x). Distributions which are not asymptotically normal may be tentatively grouped as follows: (1) Those for which $\bar{P}_n(x)$ or $\bar{P}_n(x)[1 - \bar{P}_n(x)]$ show some irregularity[3]; or as in the case of Poisson's law, distributions for which the variance σ^2 remains finite. (2) $f\{V(x)\}$ may be such that its first derivative at $\bar{P}_n(x)$ is zero. We have seen [Eqs. (21) and (21′)] that for the statistical function ω_n^2, $f'\{\bar{P}_n(x), y\} \equiv 0$. But its second derivative at this point is different from zero. The same is true of Pearson's X^2. These are statistical functions "of second class." In the final two sections some comments which generalize this case (2) will lead to a general classification of statistical functions.[4]

Before starting this more general study we may try to compare our approach and results with certain others. Certainly many of the particular results contained in our first general theorem, as well as in our laws of large numbers are well known today. What is specific to our approach and how does it compare with other approaches and results?

Historically, it should be remembered that in 1935-1936 [see v. Mises (A), (B), (C)] the study of asymptotic distributions of statistics not covered by the central limit theorem and its ramifications was hardly begun and v. Mises' original interest was as much in results as in methods. Regarding his method he says [p. 106 of (B)] "I did not endeavour to reduce the assumptions to a minimum. However, in the concept of statistical functions $f\{S_n(x)\}$, etc. I see a certain progress of probability

[3] Note that the conditions listed in Theorem B, in particular Eqs. (105) and (107) are sufficient conditions; hence their violation does not necessarily mean that the theorem does not hold.

[4] W. Hoeffding informs me of a paper by A. A. Filippowa, "Mises' theorem of the asymptotic behavior of functionals of empiric distribution functions and its statistical applications." (Russian). *Teor. Veroyatnost. i Primenen* **7** (1962), pp. 26-60.

calculus similar to the one achieved by the introduction of the general concept of distribution which I proposed in 1919."

The last decades have brought much factual and methodological progress in the study of asymptotic distributions of statistics. It would be difficult—and beyond the scope of this book—to describe the various lines of this work systematically, in a meaningful and fairly complete manner. Out of a great number of contributions, we mention a basic paper by H. B. Mann and A. Wald followed up by H. Chernoff and J. Pratt.[5] The Mann-Wald starting point was the desire to unify (and thus generalize) a manifold of particular results.

One essential line in this direction is the systematic study of the generalization of relations like $o(1)$, $O(1)$, $o(f_n)$, $O(f_n)$, to corresponding order relations on convergence "in probability" $o_p(f_n)$, $O_p(f_n)$ etc., called stochastic order relations.

We remind the reader of some definitions and add a few: let $x_1, x_2, \ldots$ be random variables. We know: x_n converges in probability to a, i.e., $x_n \underset{p}{\rightarrow} a$, if for all $\eta > 0$, $\Pr\{| x_n - a | > \eta\} \rightarrow 0$ as $n \rightarrow \infty$.

The following are likewise well known:

(a) $x_n = o(f_n)$ if $| x_n/f_n | \rightarrow 0$.

(b) $x_n = O(f_n)$ if $| x_n/f_n |$ is bounded.

Analogously we define

(a′) $x_n = o_p(f_n)$ if $| x_n/f_n | \underset{p}{\rightarrow} 0$.

(b′) $x_n = O_p(1)$: x_n is bounded in probability if to every $\epsilon > 0$ an M and an N exist such that $\Pr\{| x_n | \leqslant M\} \geqslant 1 - \epsilon$ for all $n > N$.

(b) $x_n = O_p(f_n)$ if $x_n/f_n = O_p(1)$.

The aim is to show that, essentially, the algebra of o and O extends to o_p and O_p. Later work in this direction is due to Chernoff and Pratt. The latter further generalized the concept of "in probability" and developed the parallel between stochastic and non-stochastic convergence so that in many respects one can reason with the former as if they were non-stochastic. One may state, for example, that a countable set of stochastic limit relations implies a further stochastic limit relation if the same implication holds for the corresponding non-stochastic limit relations. A well-known result following from this approach is a theorem by Slutsky (quoted in Cramér [4], p. 255 and in Mann-Wald).

[5] H. B. Mann and A. Wald, "On stochastic limit and order relationships." *Ann. Math. Statist.* **14** (1943), pp. 217-226.

H. Chernoff, "Large-sample theory: Parametric case." *Ann. Math. Statist.* **27** (1956), pp. 1-21.

J. W. Pratt, "On a general concept of "in probability." *Ann. Math. Statist.* **30** (1959), pp. 549-558.

A second line of approach attempts to establish limit distributions of one sequence of variates using known limit distributions of another. One might view our lemma of Section 5.2. in this way. A typical theorem of this kind relates to generalization and precision of the statement that if x_n has a limiting distribution then for a continuous function g, $g(x)_n$ has the corresponding limiting distribution (Chernoff's theorem 2). Briefly: if $x_n \xrightarrow{d} x$ (c.d.f. of $x_n \to$ c.d.f. of x) then $g(x_n) \xrightarrow{d} g(x)$ under conditions on g. Of course, one has to know or to find the distribution of $g(x)$. One can consider under this aspect Doob's "heuristic" proof of the distribution of ω^2. (See Chapter IX, Section 7.3.)

If we take the above description (of the "second line of approach") in its most general sense, then this is also the line of v. Mises' work. Indeed, his theorems concerning statistical functions of first class (pp. 647 and 651 of this chapter) are based on the central limit theorem, i.e., on the asymptotic distribution of the "linear" statistical functions. Also, a much more general result, Theorem I, p. 664, states that the limit distribution of "statistical functions of class r" is equal to the limit distribution of the simplest function of this type, a "polynomial of degree r" [Eq. (143)]. The problem (which v. Mises solved for $r = 1$ and $r = 2$) remains to find for given r (e.g., for $r = 3$) the just-named typical distribution.

An advantage of the Mann-Wald approach is that ready-made "cut and dried" methods of reasoning are presented in a convenient notation, leading to results as well as to methodological unification. Of course many of the results were known before or can easily be derived directly without the unifying theory. Likewise, doubtlessly, some of the results of the Mises theory follow from results of the Mann-Wald approach or by direct investigation. Mises' theory does not yet present the results as "cut and dried" as seems desirable. On the other hand, the method leads to a deeper understanding in that it concerns arbitrary functions of distribution functions (sample d.f.'s and probability d.f.'s) and leads to a complete classification of statistical functions and their distributions (see remaining part of this chapter). A further formalization of Mises' method (making use of modern theory of functionals) would be very desirable and would also make comparison more easy.

D. Classification of Differentiable Statistical Functions (Sections 7 and 8)

7. Asymptotic Expressions for Expectations

7.1. *The problem.* We introduced at the beginning of this chapter the linear, bilinear, ... statistical functions [see Eqs. (2), (4), and (5)] which correspond to homogeneous polynomials of first, second, ... degree. As in Eq. (17) we now replace $V(x)$ by $T_n(x)$ where, as in Section 6,

$$T_n(x) = S_n(x) - \bar{P}_n(x), \tag{119}$$

with $\bar{P}_n(x) = (1/n) \sum_{\nu=1}^{n} P_\nu(x)$. In the expressions

$$f_1\{S_n\} = \int a(x)\, dT_n(x), \qquad f_2\{S_n\} = \iint b(x, y)\, dT_n(x)\, dT_n(y), \ldots \tag{120}$$

$S_n(x)$ is the independent variable while $\bar{P}_n(x)$ plays the role of a "constant" [equal to $V_0(x)$ in Eq. (17)] and (120) correspond to inhomogeneous polynomials of first or second degree. The linear one is

$$\frac{1}{n}\,[a(x_1) + a(x_2) + \cdots + a(x_n)] - \int a(x)\, d\bar{P}_n(x).$$

The inhomogeneous polynomial of degree r, the *quantic of* rth *order*, in the terminology introduced by v. Mises (D), is defined as in (17), but with the present definition (119) of $T_n(x)$. It reads

$$f_r\{S_n\} = \int \cdots \int k(x_1, x_2, \ldots x_r)\, dT_n(x_1)\, dT_n(x_2) \cdots dT_n(x_r). \tag{121}$$

We shall show in the last two sections that these special statistical functions, the quantics of order $r = 1, 2, \ldots$, are, in a sense, if certain irregularities are excluded, typical statistical functions each representative of a whole class of such functions insofar as their asymptotic behavior is concerned. To show this we shall need the expectations of the quantics, and, in particular, the asymptotic values of these expectations. Every step will be exactly explained, proof and result; but the proofs will not always be complete in every detail. We begin again with the consideration of arithmetical distributions.

7.2. *Arithmetic distributions.* As in Section 5.3. we call $c_1, c_2, \ldots, c_k$ the abscissas of the discontinuities of $P_\nu(x)$; the $p_{\nu\kappa}$, $\bar{p}_\kappa$, and r_κ, $\kappa = 1, 2, \ldots, k$ have the same meaning as in Section 5, i.e., $p_{\nu\kappa}$ is the step of $P_\nu(x)$ at c_κ, $\bar{p}_\kappa = (1/n) \sum_{\nu=1}^{n} p_{\nu\kappa}$, and r_κ is the observed relative frequency of c_κ in n observations; let also $t_\kappa = r_\kappa - \bar{p}_\kappa$. Then (121) becomes

$$f_r\{S_n(x)\} = \sum_{\alpha} k(c_{\alpha_1}, c_{\alpha_2}, \ldots, c_{\alpha_r}) t_{\alpha_1} t_{\alpha_2}, \ldots, t_{\alpha_r}, \tag{122}$$

where the $\alpha_1, \alpha_2, \ldots, \alpha_r$ take on the values 1, 2, ..., k. We wish to compute the expectation of f_r, keeping in mind that $k(c_{\alpha_1}, \ldots, c_{\alpha_r})$ is, in (122), a constant coefficient which could be written $k_{\alpha_1, \alpha_2, \ldots, \alpha_r}$, and we have to compute the expectation of

$$t_{\alpha_1}, t_{\alpha_2}, \ldots, t_{\alpha_r} \tag{123}$$

only.

Now, note that the present t_κ are, except for the factor $1/n$, nothing but $\delta_\kappa = n_\kappa - nq_\kappa$, whose expectations we computed in determining the variance of Pearson's X^2, of ω^2, etc. (see Chapter IX, Section 4). From the Bernoulli distribution (multinomial distribution for $k = 2$) we had (writing now p instead of the q used in Chapter IX)

$$E_n[\delta_1] = 0, \qquad E_n[\delta_1^2] = np_1(1 - p_1), \qquad E_n[\delta_1^3] = np_1(1 - p_1)(2p_1 - 1)$$
$$E_n[\delta_1^4] = 3n^2p_1^2(1 - p_1)^2 + np_1(1 - p_1)[1 - 6p_1(1 - p_1)].$$

A general result (see Chapter IV, Section 3) was that the central moments of order $2m$ and of order $(2m + 1)$ of the Bernoulli distribution are polynomials in n of degree m.

The moments of the multinomial distribution with $k = 3$ are also well known. We had in Chapter IX

$$E_n[\delta_1\delta_2] = -np_1p_2, \qquad E_n[\delta_1^2\delta_2] = np_1p_2(2p_1 - 1)$$
$$E_n[\delta_1^2\delta_2^2] = np_1p_2[(n - 2)(1 - p_1 - p_2 + 3p_1p_2) + 1], \quad \text{etc.}$$

In general, we need the moments of the r-variate multinomial distribution, or actually the term of the highest degree in n only—an essentially elementary problem. The method to be used is that of characteristic functions which gave also the result in Chapter XI, Section 8 when we estimated the expectations of $\delta_1\delta_2\delta_3$ and $\delta_1^2\delta_2^2\delta_3^2$. The main results, of which all the preceding formulas are particular cases, are as follows [for complete proofs see v. Mises (D), Part I]:

The expectation $E_n[\delta_1^\alpha\delta_2^\beta\delta_2^\gamma \cdots]$, $\alpha + \beta + \gamma + \ldots = r = 2m$ *is a polynomial in* n *of order* $m = r/2$; *and if* $\alpha + \beta + \gamma + \cdots = r = 2m + 1$, *it is likewise a polynomial of order* m, *hence, of an order less than* $r/2$.

Since $t_\kappa = (1/n)\,\delta_\kappa$ it is seen that the expectation of $t_1^\alpha t_2^\beta t_3^\gamma \cdots$, $\alpha + \beta + \gamma + \cdots = 2m$ is of order $n^m/n^{2m} = n^{-m}$, and if $\alpha + \beta + \gamma + \cdots = 2m + 1$, it is of the order $n^m/n^{2m+1} = n^{-(m+1)}$. Hence, for the f_r of (121) the result

$$\lim_{n\to\infty} n^{r/2}E_n[f_r] = 0 \qquad r \text{ odd}$$
$$\neq 0 \qquad r \text{ even.} \tag{124}$$

The Schwarz' inequality applied to any function C of x_1, x_2, ..., x_n reads

$$E_n[|C|] \leqslant \sqrt{E_n[C^2]}. \tag{125}$$

Therefore, for any $\delta > 0$ and any r, odd or even,

$$\lim_{n\to\infty} n^{(r/2)-\delta} E_n[|f_r|] = 0. \tag{126}$$

Next, we have to find for $r = 2m$ the value of

$$\lim_{n\to\infty} n^m E_n[t_1{}^\alpha t_2{}^\beta \cdots], \qquad \alpha + \beta + \cdots = 2m. \tag{127}$$

The result will, of course, depend on the given p_ν. The interesting answer is the following: The value of (127) depends on two kinds of magnitudes only which form a "matrix of variances." Put

$$\begin{aligned} S_{ii}^{(\nu)} &= p_{\nu i}(1 - p_{\nu i}) \\ S_{i\kappa}^{(\nu)} &= -p_{\nu i} p_{\nu\kappa} \end{aligned} \tag{128}$$

or, with the Kronecker delta $\delta_{i\kappa}$, where $\delta_{ii} = 1$, $\delta_{i\kappa} = 0$:

$$S_{i\kappa}^{(\nu)} = \delta_{i\kappa} p_{\nu i} - p_{\nu i} p_{\nu\kappa}. \tag{128'}$$

From these we form $\bar{S}_{ii} = (1/n) \sum_{\nu=1}^{n} S_{ii}^{(\nu)}$, $\bar{S}_{i\kappa} = (1/n) \sum_{\nu=1}^{n} S_{i\kappa}^{(\nu)}$. If we introduce $(1/n) \sum_{\nu=1}^{n} p_{\nu i} p_{\nu\kappa} = \overline{p_i p_\kappa}$, we have

$$\begin{aligned} \bar{S}_{ii} &= \bar{p}_i - \overline{p_i p_i} \\ \bar{S}_{i\kappa} &= -\overline{p_i p_\kappa}. \end{aligned} \tag{129}$$

These are the decisive magnitudes. Note that the $\bar{S}_{i\kappa}$ are symmetric.

With this notation we have the final result,[1] *if* $r = 2m = \alpha + \beta + \gamma + \cdots$, and $Q = \frac{1}{2} \sum_{i,\kappa} \bar{S}_{i\kappa} u_i u_\kappa$, then

$$\lim_{n\to\infty} n^m E_n[t_1{}^\alpha t_2{}^\beta t_3{}^\gamma \cdots] = \frac{1}{m!} \frac{\partial^{2m}}{\partial u_1{}^\alpha \partial u_2{}^\beta \partial u_3{}^\gamma \cdots} \left[\left(\frac{1}{2} \sum_{i,\kappa} \bar{S}_{i\kappa} u_i u_\kappa\right)^m\right]. \tag{130}$$

This result can also be stated as follows. Denote the right-hand side of

[1] All these results which are properties of the multinomial distribution are elementary; the detailed derivation, which is lengthy, is in v. Mises (D), pp. 314-318; see also (E), Sections 14 and 15.

(130) by $C_{\alpha,\beta,\gamma,\ldots}$. Then, Eq. (130) states that $C_{\alpha,\beta,\gamma,\ldots}$ are derived as coefficients in the expansion of the mth power of the quadric Q, namely:

$$\left(\frac{1}{2}\sum_{i,\kappa} \bar{S}_{i\kappa}t_i t_\kappa\right)^m = m!\sum \frac{C_{\alpha,\beta,\gamma,\ldots}}{\alpha!\,\beta!\,\gamma!\,\cdots}\, t_1{}^\alpha t_2{}^\beta t_3{}^\gamma \cdots . \tag{130'}$$

To exemplify our last results take $r = 2m = 4$, $\alpha = 2$, $\beta = 2$; hence, compute the asymptotic expectation of $(1/n^2)\,\delta_1{}^2\delta_2{}^2 = t_1{}^2t_2{}^2$. According to (130) we form

$$[\tfrac{1}{2}(\bar{S}_{11}u_1{}^2 + 2\bar{S}_{12}u_1u_2 + \bar{S}_{22}u_2{}^2)]^2 = \tfrac{1}{4}[4\bar{S}_{12}^2u_1{}^2u_2{}^2 + 2\bar{S}_{11}\bar{S}_{22}u_1{}^2u_2{}^2 + \cdots].$$

Since we have to compute the fourth mixed derivative the differentiation of terms other than $u_1{}^2u_2{}^2$ will give zero. Therefore, we have

$$\frac{1}{2!}\,\frac{\partial^4}{\partial u_1{}^2\,\partial u_2{}^2}\,[\bar{S}_{12}^2u_1{}^2u_2{}^2 + \tfrac{1}{2}\bar{S}_{11}\bar{S}_{22}u_1{}^2u_2{}^2] = 2\bar{S}_{12}^2 + \bar{S}_{11}\bar{S}_{22}\,. \tag{131}$$

If all P_ν are equal this reduces to $p_1p_2(1 - p_1 - p_2 + 3p_1p_2)$ as in the previous formula for $E[\delta_1{}^2\delta_2{}^2]$.

As a second example, we find in the same way:

$$\lim_{n\to\infty} n^2E_n[t_1t_2t_3t_4] = \bar{S}_{12}\bar{S}_{34} + \bar{S}_{13}\bar{S}_{24} + \bar{S}_{14}\bar{S}_{23}\,. \tag{132}$$

The corresponding result for general (even) r is

$$\lim_{n\to\infty} n^{r/2}E_n[t_1t_2\cdots t_r] = \sum \bar{S}_{i_1\kappa_1}\bar{S}_{i_2\kappa_2}\cdots\bar{S}_{i_r\kappa_r}\,, \tag{132'}$$

where $i_1, \kappa_1, \ldots, i_r, \kappa_r$ are any arrangement of 1, 2, ..., r. For $r = 6$, for example, there are 15 terms corresponding to the subscripts:

12 | 34 | 56; 12 | 35 | 46; 12 | 36 | 45; 13 | 24 | 56; 13 | 25 | 46;

13 | 26 | 45; 14 | 23 | 56; 14 | 25 | 36; 14 | 26 | 35; 15 | 23 | 46;

15 | 24 | 36; 15 | 26 | 34; 16 | 23 | 45; 16 | 24 | 35; 16 | 25 | 34.

The result (132′) is in a way general since it remains correct regardless of whether the subscripts are or are not different from one other.

If we want, for example the $\lim_{n\to\infty} n^3E_n[t_1{}^2t_2{}^2t_3{}^2]$ we put in the above 15 terms $4 = 1$, $5 = 2$, $6 = 3$ and obtain

$$\lim_{n\to\infty} n^3E_n[t_1{}^2t_2{}^2t_3{}^2] = \bar{S}_{11}\bar{S}_{22}\bar{S}_{33} + 2(\bar{S}_{11}\bar{S}_{23}^2 + \bar{S}_{22}\bar{S}_{13}^2 + \bar{S}_{33}\bar{S}_{12}^2) + 8\bar{S}_{12}\bar{S}_{23}\bar{S}_{13},$$

or likewise

$$\lim_{n\to\infty} n^3 E_n[t_1{}^4 t_2{}^2] = 3\bar{S}_{11}^2 \bar{S}_{22} + 12\bar{S}_{11}\bar{S}_{12}^2$$

$$\lim_{n\to\infty} n^3 E_n[t_1{}^6] = 15\bar{S}_{11}^3 .$$

The number of terms in (132′) is

$$\frac{(2m)!}{2^m m!} = 1 \cdot 3 \cdot 5 \cdots (2m-1), \tag{133}$$

each term counted as often as it appears. Since in the polynomial $f_r = f_{2m}$ the $k(x_1, x_2, ..., x_{2m})$ is symmetric, it follows that *the asymptotic expectation of* f_{2m} may be written

$$\begin{aligned} &\lim_{n\to\infty} n^m E_n[f_{2m}] = \\ &1 \cdot 3 \cdot 5 \cdots (2m-1) \sum_{(\alpha)} k(c_{\alpha_1} c_{\alpha_2} \cdots c_{\alpha_{2m}}) \cdot \bar{S}_{\alpha_1\alpha_2} \bar{S}_{\alpha_3\alpha_4} \cdots \bar{S}_{\alpha_{2m-1}\alpha_{2m}} . \end{aligned} \tag{134}$$

7.3. *Continuous distributions.* There is no difficulty transferring the results of Section 7.2 to continuous distributions as long as the $P_\nu(x)$ are uniformly bounded.

Definition (122) is then replaced by

$$f_r\{S_n(x)\} = \int \cdots \int k(x_1, x_2, ..., x_r)\, dT_n(x_1)\, dT_n(x_2) \cdots dT_n(x_r).$$

The entire computation can be repeated in terms of the $dT_n(x_1)$, $dT_n(x_2)$, ... instead of the $t_1, t_2, ...$ (see Chapter IX, Section 6.2, the computation of the variance of ω^2). For the increment $p_{\nu i}$ we now have $dP_\nu(x)$, and $\delta(x, y)$, where $\delta(x, x) = 1$, $\delta(x, y) = 0$, $x \neq y$, takes the place of the Kronecker symbol $\delta_{\iota\kappa}$. The place of the $S_{\iota\kappa}^{(\nu)}$ is taken now (see again Chapter IX, Section 6.2) by

$$\begin{aligned} B_\nu(x, y) &= P_\nu(x) - P_\nu(x)P_\nu(y), \qquad x \leqslant y \\ &= P_\nu(y) - P_\nu(x)P_\nu(y), \qquad x \geqslant y. \end{aligned} \tag{135}$$

Each $B_\nu(x, y)$ is the difference between two two-dimensional distributions. One is different from zero along the line $x = y$ and its linear density is $P_\nu'(x)/\sqrt{2}$, [the prime in $P_\nu'(x)$ means differentiation] the other has the density $P_\nu'(x)\, P_\nu'(y)$ in the whole x, y-plane.

The arithmetical means $\bar{S}_{i\kappa}$ are to be replaced by

$$\begin{aligned} \bar{B}(x, y) &= \bar{P}_n(x) - \frac{1}{n}\sum_{\nu=1}^{n} P_\nu(x)P_\nu(y), \qquad x \leqslant y \\ &= \bar{P}_n(y) - \frac{1}{n}\sum_{\nu=1}^{n} P_\nu(x)P_\nu(y), \qquad x \geqslant y. \end{aligned} \tag{136}$$

The formula (134) is replaced by

$$\lim n^m E_n[f_{2m}] = \\ 1 \cdot 3 \cdots (2m-1) \int\int\cdots\int k(x_1, x_2, ..., x_{2m})\, d\bar{B}(x_1, x_2)\, d\bar{B}(x_3, x_4) \cdots d\bar{B}(x_{2m-1}, x_{2m}). \tag{137}$$

We introduce into (137) the definitions (136) and the abbreviation $\bar{\bar{P}}_n(x, y) = (1/n) \sum_{\nu=1}^{n} P_\nu(x)P_\nu(y)$. The right-hand side of (137) is then a sum of integrals of multiplicities ranging from m to $2m$. We have, for example, for $m = 2$, $r = 4$:

$$\begin{aligned} \lim_{n\to\infty} n^2 E_n[f_4] = 1 \cdot 3 \Big[&\int\int k(x_1, x_1, x_2, x_2)\, d\bar{P}_n(x_1)\, d\bar{P}_n(x_2) \\ &- 2 \int\int\int k(x_1, x_1, x_3, x_4)\, d\bar{P}_n(x_1)\, d\bar{\bar{P}}_n(x_3, x_4) \\ &+ \int\int\int\int k(x_1, x_2, x_3, x_4)\, d\bar{\bar{P}}_n(x_1, x_2)\, d\bar{\bar{P}}_n(x_3, x_4) \Big]. \end{aligned} \tag{138}$$

As remarked at the beginning, these results hold if the $P_\nu(x)$ are uniformly bounded. In the general case we need some restriction regarding the behavior of $P_\nu(x)$ as $x \to \infty$. The following conditions are sufficient for the convergence of the integrals in (137): Given a sequence of distributions $P_1(x)$, $P_2(x)$, ... assume that $k(x_1, x_2, ..., x_r)$ (of (137)) is the sum of several terms for each of which *there exists a continuous* majorant of the form

$$\alpha(x_1)\alpha(x_2) \cdots \alpha(x_r), \tag{139}$$

where the integrals $\int \alpha^2(x)\, dP_\nu(x)$ *have a finite upper bound independent of* n.[2] If we glance at (138) and remember the meaning of the $\bar{P}_n(x)$ and $\bar{\bar{P}}_n(x, y)$ we see that under the assumed condition the integral in (137) has a finite value. It follows that (126) remains true in the general case.

[2] See v. Mises (E), p. 22; (D), p. 321.

8. Asymptotic Behavior of Statistical Functions

8.1. *The statements.* We consider again an infinite sequence of probability distributions $P_1(x)$, $P_2(x)$, ... and a differentiable statistical function $f\{V(x)\}$. Let $x_1, x_2, ..., x_n$ be the results of n observations where $P_\nu(x)$ is the c.d.f. for the νth trial, $\nu = 1, 2, ... n$ and $S_n(x)$ the repartition of the $x_1, x_2, ..., x_n$. Our aim is a statement on the asymptotic distribution of $f\{S_n(x)\}$. In line with Sections 6 and 7 we consider the difference $f\{S_n(x)\} - f\{\bar{P}_n(x)\}$, where $\bar{P}_n(x)$ is the arithmetical mean of the $P_\nu(x)$, $\nu = 1, 2, ..., n$. With an r, which has the same meaning as in Section 7 and was equal to unity in Sections 5 and 6, we put, as in (98),

$$A_n = n^{r/2}[f\{S_n(x)\} - f\{\bar{P}_n(x)\}]. \tag{140}$$

This will be the variate whose asymptotic distribution we wish to find.

We make now the following assumptions:

(a) *The function $f\{V(x)\}$ has derivatives up to the order $(r + 1)$, $r = 1, 2, ...,$ in a convex set J of distributions which contains the $\bar{P}_n(x)$ at least from a certain n_0 on and which contains all $S_n(x)$ that can occur.*

(b) *The first $(r - 1)$ derivatives of $f\{V(x)\}$ vanish at $\bar{P}_n(x)$.*

(c) *The* rth *derivative of $f\{V(x)\}$ at $\bar{P}_n(x)$ is equal to a function $k_n(x_1, x_2, ..., x_r)$ which, with respect to the given $P_\nu(x)$, satisfies the conditions stated at the end of Section 7.*

(d) *The* $(r + 1)$th *derivative of $f\{V(x)\}$ is bounded for all V in J in the sense of condition* (24'), (p. 630) *with a $\psi(x)$ such that $\int \Psi(x)\, dP_\nu(x)$ is bounded, where $\Psi(x)$ is the indefinite integral of $\psi(x)$.*

We shall prove:

THEOREM I. *Denote by B_n the statistical function*

$$B_n = n^{r/2} \frac{1}{r!} \int \cdots \int k_n(x_1, x_2, ..., x_r)\, dT_n(x_1)\, dT_n(x_2) \cdots dT_n(x_r), \tag{141}$$

a "quantic" of order r, multiplied by $n^{r/2}$.

The asymptotic distribution of the A_n of (140) *is equal to that of B_n for those x where the asymptotic distribution of B_n* is continuous.

THEOREM II. *Under the conditions of Theorem I, the moments of any order of the asymptotic distribution of A_n depend* [as far as the $P_\nu(x)$ are concerned] *only on the means*

$$\bar{P}_n(x) = \frac{1}{n}\sum_{\nu=1}^{n} P_\nu(x) \qquad \text{and} \qquad \bar{\bar{P}}_n(x, y) = \frac{1}{n}\sum_{\nu=1}^{n} P_\nu(x)P_\nu(y)\,. \tag{142}$$

8.2. *The proofs.* We apply Taylor's formula (29) to $f\{S_n(x)\} - f\{\bar{P}_n(x)\}$ and obtain, on account of assumption (b)

$$f\{S_n(x)\} - f\{\bar{P}_n(x)\} = \frac{1}{r!}\int \cdots \int f^{(r)}\{\bar{P}_n(x), \xi_1, \xi_2, \dots, \xi_r\}\, dT_n(\xi_1) \cdots dT_n(\xi_r)$$
$$+ \frac{1}{(r+1)!}\int \cdots \int f^{(r+1)}\{P(x), \xi_1, \xi_2, \dots, \xi_{r+1}\}\, dT_n(\xi_1) \cdots dT_n(\xi_{r+1}). \tag{143}$$

We multiply both sides of this equation by $n^{r/2}$, introduce the A_n and B_n of (140) and (141), and obtain

$$A_n - B_n = \frac{n^{r/2}}{(r+1)!}\int \cdots \int f^{(r+1)}\{P(x), \xi_1, \dots, \xi_{r+1}\}\, dT_n(\xi_1) \cdots dT_n(\xi_{r+1}). \tag{144}$$

On account of condition (d) and condition (24′) the integral in (144) is less in absolute value than[1]

$$\left[\int T_n^2(x)\psi(x)\, dx\right]^{(r+1)/2},$$

and, therefore,

$$| A_n - B_n | \leqslant \frac{n^{r/2}}{(r+1)!}\left[\int T_n^2(x)\psi(x)\, dx\right]^{(r+1)/2}. \tag{145}$$

By integration by parts this integral may be transformed. Put

$$\begin{aligned} \beta(x, y) &= \int_0^x \psi(z)\, dz = \Psi(x) \qquad \text{for } x \leqslant y \\ &= \int_0^y \psi(z)\, dz = \Psi(y) \qquad \text{for } x \geqslant y. \end{aligned} \tag{146}$$

Then

$$I = \int \psi(x) T_n^2(x)\, dx = \int\int \beta(x, y)\, dT_n(x)\, dT_n(y) = f_2\,. \tag{147}$$

The $(r+1)$th power of f_2 is an f_{2r+2}, non-negative and dependent on $\psi(x)$. The "coefficients" of f_{2r+2} which take the place of the $k(\cdots)$ of (137) are products $\beta(x_1, x_2)\beta(x_3, x_4), \dots$; this f_{2r+2} will exist if $\int \Psi(x)\, dP_\nu(x)$ is bounded, as assumed in (d).

We have from (145), using (147):

$$| A_n - B_n |^2 \leqslant \frac{n^r}{[(r+1)!]^2} f_{2r+2}\,, \tag{148}$$

[1] Compare (105) and (106). From here on we need a different proof.

and from (126)

$$\lim_{n\to\infty} n^r E_n[f_{2r+2}] = 0. \tag{149}$$

Therefore, by (125) and by (148)

$$\lim_{n\to\infty} E_n[| A_n - B_n |] = 0. \tag{150}$$

We apply now the lemma of Section 5.2, and see that the asymptotic distributions of A_n and B_n coincide as stated in Theorem I.

As for Theorem II, we know that if a sequence of distributions tends toward a limit distribution, then the moments of any order of the sequence tend toward the respective ones of the limit distribution. All moments of B_n, which is an f_r, are expectations of certain polynomials[2] and the general form of these expectations is given in (137). We have seen before that under the conditions formulated at the end of Section 7, the integrals in (137) exist and Theorem II is thus proved.

We can give it the following form:

THEOREM II′. *Under the conditions of Theorem I, the asymptotic distribution of $f\{S_n(x)\}$ is "essentially determined" by the first non-vanishing derivative of $f\{V(x)\}$ at the point $\bar{P}_n(x)$ and by the averages* (142).

By "essentially determined" is meant, determined except for an additional function whose moments of any order are zero. An example of such a function may be found, for example in Shohat and Tamarkin, *The Problem of Moments*, New York, 1943; it is reproduced in v. Mises (D, p. 329).[3] For statistical purposes, it seems natural to consider two distributions as "essentially the same", if they have the same moments.

8.3. Classification of statistical functions. Our Theorems I and II show that if certain regularity assumptions hold regarding the $\bar{P}_n(x)$, $\bar{\bar{P}}_n(x, y)$ and the statistical functions, *there exists a sequence of asymptotic distributions, of which the Gaussian is the first, dependent on the order* $r = 1, 2, \ldots$ *of the first non-vanishing derivative of $f\{V(x)\}$ at $\bar{P}_n(x)$.* We can find all types of statistical functions (disregarding "exceptional" behavior[4]) by investigating the asymptotic behavior of the quantics f_r

[2] See details on computations of moments of quantics in v. MISES (D), p. 320.

[3] This does not contradict the uniqueness theorem (Chapter VIII) which states that under certain circumstances a distribution is determined by its moments in a unique way, since that theorem requires that the density of the distribution vanishes at infinity in a sufficiently strong manner.

[4] It would be important to study systematically the cases of "exceptional behavior," mentioned also in Section 6.4.

given in (121). The functions of *type one* are those whose first derivative is non-zero at $V(x) = \bar{P}_n(x)$ and their asymptotic distribution is the normal one (Sections 5 and 6).

The *type two* has been mentioned in Chapter IX. Its asymptotic distribution is similar to that of ω^2. To find the general form of the asymptotic distribution of type two we have to investigate that of

$$f\{S_n(x)\} = f_2 = \iint k(x, y)\, dT_n(x)\, dT_n(x), \qquad T_n = S_n - \bar{P}_n\,, \tag{151}$$

where $k(x, y)$ may be supposed symmetric.

If, in particular,

$$\begin{aligned} k(x, y) &= g(x), \qquad x \leqslant y \\ &= g(y), \qquad x \geqslant y, \end{aligned} \tag{152}$$

then integration by parts shows [see (146) and (147)] that

$$f\{S_n(x)\} = \int g'(x) T_n^2(x)\, dx, \tag{153}$$

where g' is the derivative of $g(x)$. The right-hand side of (153) with $\lambda(x)$ instead of $g'(x)$ is the ω^2 of Chapter IX. The complete theory of the distributions of type two is discussed in (D), pp. 331-348 and in (*E*), pp. 29-37.

We ask finally *whether types of any order exist.* The answer is rather obvious. Consider the linear statistical function $f\{S_n(x)\} = \int a(x)\, dT_n(x)$, $T_n(x) = S_n - \bar{P}_n$, whose asymptotic distribution is given by

$$\lim_{n\to\infty} \Pr\{\sqrt{n} f\{S_n(x)\} \leqslant u\} = \Phi\Big(\frac{u}{\sigma_n}\Big) \tag{154}$$

$$\sigma_n^2 = \frac{1}{n}\sum_{\nu=1}^{n} \Big[\iint a^2(y)\, dP_\nu(y) - \iint a(y) a(z)\, dP_\nu(y)\, dP_\nu(z)\Big]. \tag{154'}$$

Denote by $h\{S_n(x)\}$ the rth power of $f\{S_n(x)\}$. Equation (154) is identical with

$$\lim_{n\to\infty} \Pr\{n^{r/2} h\{S_n(x)\} \leqslant u\} = \Phi\Big(\frac{u^{1/r}}{\sigma_n}\Big). \tag{155}$$

The asymptotic density of $h = f^r$ is given by

$$\frac{1}{\sigma_n r\sqrt{2\pi}}\, u^{(1/r)-1} \exp\Big(-\frac{u^{2/r}}{\sqrt{2}\sigma_n^{\,2}}\Big).$$

For $r = 2$ we recognize the χ^2-distribution.

Since f vanishes at $\bar{P}_n(x)$ the rth power of f and all its derivatives of orders 1, 2, ..., $r - 1$ vanish at $\bar{P}_n(x)$. Hence, h is a particular example of a statistical function whose type is of order r.

This example can be generalized by taking as a starting point any statistical function $f\{V(x)\}$ of known asymptotic distribution and using for h an (ordinary) function of f whose Taylor expansion at $f = 0$ begins with the term f^r.