CHAPTER III

BASIC PROPERTIES OF DISTRIBUTIONS

A. Mean Value, Variance, and Other Moments (Sections 1-4)

After the more difficult considerations of Chapter II, we turn to the study of some basic concepts of an elementary nature.

In Sections 3 and 7 of Chapter II we introduced the d.f. $F(x)$. We denoted it by $F(x)$ in order to distinguish it from the probability distribution $P(A)$, which is a set function. This set function will not be used much in the remainder of this book since the point function $F(x)$ will prove to be more convenient. We shall, from now on, *denote the c.d.f. by* $P(x)$ and the corresponding density or discrete probabilities by $p(x)$ or p_i. In Chapter VIII we shall then introduce empirical distribution functions as contrasted to probability distribution functions and shall modify our notation accordingly.

1. *Mean Value and Variance. Tchebycheff's Inequality*

1.1. *Mean value, variance, and median.* Consider a collective with a one-dimensional arithmetical (= discrete) distribution the label values being $x_1, x_2, \ldots, x_k$. Suppose the ith label x_i occurred n_i times among the first n elements of the collective. The expression

$$\frac{1}{n}\sum_{i=1}^{k} n_i x_i = x_1 \frac{n_1}{n} + x_2 \frac{n_2}{n} + \cdots + x_k \frac{n_k}{n} \tag{1}$$

is then the average label value for the first n elements. According to our definition of the probabilities p_i, this expression, for ever increasing n, will tend to the limit

$$a = \sum_{i=1}^{k} x_i p_i = \sum_{i=1}^{k} x_i p(x_i). \tag{2}$$

If there are countably many labels x_i the summation is from 1 to ∞. The quantity a defined by (2) is called the *mean value* of the distribution under consideration. In the case of countably many labels, the sum in (2) is infinite and we assume that it converges absolutely whenever we say the mean value exists.

In the case of a one-dimensional continuous distribution with density $p(x)$, the definition of the mean value takes the form

$$a = \int_{-\infty}^{+\infty} xp(x)\,dx, \tag{3}$$

where a exists if the integral converges.

Consider the density $p(x) = [\pi(1 + x^2)]^{-1}$ (Cauchy density) defined in $(-\infty, +\infty)$. The mean value is given by the integral

$$\frac{1}{\pi}\int \frac{x}{1+x^2}\,dx$$

which is divergent. The mean value does not exist.

Applying the mechanical analogy, it is seen that the mean value a is the abscissa of the centroid of a mass 1 distributed along the x-axis in lumps p_i or in a continuous way with the density $p(x)$.

We consider some simple examples: The mean value of *an alternative* with probabilities p and $q = 1 - p$ corresponding to the label values 0 and 1, respectively, is $0 \cdot p + 1 \cdot q = q$. The mean value of the numbers resulting from the throwing of a correct die is $\frac{1}{6}(1+2+\cdots+6) = 3.5$. As an example of the continuous case take the constant density $p(x) = 1$ for $0 \leqslant x \leqslant 1$, and $p(x) = 0$ outside that interval. Here $\int_{-\infty}^{+\infty} p(x)\,dx = \int_0^1 dx = 1$, as required, and $a = \int_{-\infty}^{+\infty} xp(x)\,dx = \int_0^1 x\,dx = \frac{1}{2}$. As another example let $p(x)$ increase linearly to one in the interval (x_1, x_2), the initial value $p(x_1)$ being zero.[1] Outside that interval, p(x) vanishes everywhere. This leads to $p(x) = c(x - x_1)$ for $x_1 \leqslant x \leqslant x_2$ where c must be determined from $c\int_{x_1}^{x_2}(x - x_1)\,dx = \frac{1}{2}c(x_2 - x_1)^2 = 1$. The mean value a is then found as

$$a = \frac{2}{(x_2 - x_1)^2}\int_{x_1}^{x_2} x(x - x_1)\,dx = \frac{2}{(x_2 - x_1)^2}\,\frac{(x_2 - x_1)^2}{6}\,(2x_2 + x_1)$$

$$= x_1 + \tfrac{2}{3}(x_2 - x_1). \tag{4}$$

The mean value is the simplest parameter that characterizes a distribution. In probability theory and particularly in the theory of

[1] In Chapter II the distinction between open, closed, and half open intervals was of importance. In the remainder of the book it rarely matters and if not otherwise, we simply write (a, b) for the interval $a \leqslant x \leqslant b$.

statistics, a great many other parameters are used to describe certain features of a distribution. We mention the *median* which may be defined as any root of the equation $P(x) = \frac{1}{2}$. In the graph of the c.d.f. $P(x)$ we draw a horizontal line at the distance $\frac{1}{2}$ from the x-axis; if it intersects $P(x)$ at one point, the abscissa of this point is the median. If the line $y = \frac{1}{2}$ intersects the step with ordinates $P(a_x)$ and $P(a_x - 0)$, respectively the median is a_x. If the horizontal line coincides with the curve of $P(x)$ along a step which, say, extends between the x-values α and β, the median is undetermined. Often the point halfway between α and β is taken as the median. We shall return to this definition in Chapter VIII, Section 2.1. In the same way as we have defined the median by means of the equation $P(x) = \frac{1}{2}$ we may define quartiles, deciles, percentiles, etc. The *lower* and *upper quartile* q_1 and q_3 are defined by $P(x) = \frac{1}{4}$ and $P(x) = \frac{3}{4}$, respectively, the deciles and percentiles in a similar way. Any of these quantities may be indeterminate in the same way as the median. The quantity $\frac{1}{2}(q_3 - q_1)$ is sometimes used as a simple measure of dispersion and is called the *semi-interquartile* range. (See more on these measures in Chapter VIII, Section 2.1.)

If we think of the distribution as a mass distribution, the next concept of interest is a quantity significant of the degree of concentration of the mass around its centroid. It is most usual to define as a measure of concentration a quantity called the *variance*, which is in close analogy to the moment of inertia:

$$s^2 = \sum_{i=1}^{k} (x_i - a)^2 p(x_i) \qquad \text{(arithmetical case)} \tag{5}$$

$$s^2 = \int_{-\infty}^{+\infty} (x - a)^2 p(x)\, dx \qquad \text{(continuous case)} \tag{6}$$

where the notation s^2 is justified by the fact[2] that the sum (integral) can never be negative. If in (5) there is an infinite sum or if in (6) the region where $p(x) \neq 0$ extends to infinity, the existence of the sum (integral) must be checked. In the continuous case, s^2 is necessarily positive. The only case in which $s^2 = 0$ is that of an arithmetical distribution with one single jump of magnitude 1, whose abscissa, then, must be equal to a. This is the case of extreme concentration; in general,

[2] The use of latin letters a and s^2 in these definitions has nothing to do with "sample distribution," "sample mean," etc. As long as we do not contrast theoretical and empirical distributions (Chapter VIII, Section 1.2) we use latin letters for convenience.

If in an integral the lower limit is $-\infty$ and the upper is $+\infty$ we shall often omit these limits.

the larger the value of s^2, the wider the dispersion. Instead of (5) or (6), we may wish to consider the dispersion around a point other than a, say x_0:

$$s_0^2 = \sum_{i=1}^{k} (x_i - x_0)^2 p(x_i) \tag{5'}$$

$$s_0^2 = \int (x - x_0)^2 p(x)\, dx. \tag{6'}$$

If we write $x - x_0 = (x - a) - (x_0 - a)$ and expand the square, we obtain, say in (6')—and quite analogously for (5')—

$$s_0^2 = \int (x - x_0)^2 p(x)\, dx = \int (x - a)^2 p(x)\, dx - 2(x_0 - a) \int (x - a) p(x)\, dx$$
$$+ (x_0 - a)^2 \int p(x)\, dx = s^2 + 0 + (x_0 - a)^2.$$

Hence, the so-called *shift-of-origin rule*

$$s_0^2 = s^2 + (x_0 - a)^2, \tag{7}$$

and in particular for $x_0 = 0$ the much used formula

$$s^2 = \int x^2 p(x)\, dx - a^2. \tag{7'}$$

Equation (7) shows that the dispersion with respect to the mean value is smaller than that with respect to any other point, x_0. Conversely, this minimal character of s_0^2 for $x_0 = a$ can serve as a definition of a.

The positive root s of the variance is called the *standard deviation.*

The variance of a uniform arithmetical distribution with label values 1, 2, ..., k is found as follows: From $p(x_i) = 1/k$ and $a = (1/k) \sum_{i=1}^{k} i = \frac{1}{2}(k + 1)$ we conclude

$$s^2 = \frac{1}{k} \sum_{i=1}^{k} \left(i - \frac{k+1}{2} \right)^2 = \frac{1}{k} \sum_{i=1}^{k} i^2 - \frac{2}{k} \frac{k+1}{2} \sum_{i=1}^{k} i + \frac{k}{k} \left(\frac{k+1}{2} \right)^2 = \frac{k^2 - 1}{12},$$

where the formulas $\sum_{i=1}^{k} i = \frac{1}{2}k(k + 1)$ and $\sum_{i=1}^{k} i^2 = \frac{1}{6}k(k + 1)(2k + 1)$ have been used. The variance of the uniform geometric distribution with $p(x) = 1$ in (0, 1) equals $\frac{1}{12}$. More on mean and variance will be found in the two following sections.

1.2. *Tchebycheff's inequality.* If the mean value and variance of a distribution are known, the distribution is by no means determined.

However, an inequality for the difference $P(a + X) - P(a - X)$, where X is any positive number greater than s, can be derived from the knowledge of a and s^2. In other words, the probability for the chance variable x to fall inside the neighborhood $(a - X, a + X)$ of the mean value a has a definite lower bound. We prove this result first for a continuous distribution $p(x)$.

Call A the open interval $(a - X < x < a + X)$ and A' the complementary region consisting of the points $x \leqslant a - X$ and $x \geqslant a + X$ (Fig. 2). Then

$$s^2 = \int_{A'} (x - a)^2 p(x)\, dx + \int_{A} (x - a)^2 p(x)\, dx \geqslant \int_{A'} (x - a)^2 p(x)\, dx.$$

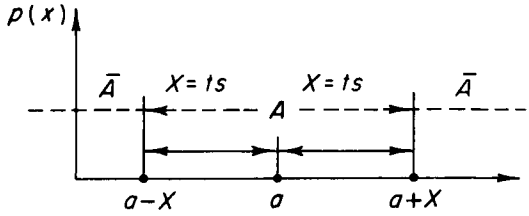

FIG. 2.

Since $(x - a)^2$ and $p(x)$ are non-negative everywhere and the minimum of $(x - a)^2$ in the region A' is X^2, we have

$$s^2 \geqslant X^2 \int_{A'} p(x)\, dx = X^2 P(A') = X^2(1 - P(A)), \tag{8}$$

with $P(A')$ and $P(A)$ here denoting the probability that x falls in the regions A' and A, respectively. By solving (8) for $P(A)$ we arrive at *Tchebycheff's inequality*:

$$P(A) \geqslant 1 - \frac{s^2}{X^2}. \tag{9}$$

The content of (9) becomes trivial if $X = s$.

Instead of $P(A)$ we may write $\Pr\{x \in A\}$ or $\Pr\{|\, x - a\,| < X\}$ and obtain

$$\Pr\{|\, x - a\,| < X\} \geqslant 1 - \frac{s^2}{X^2}.$$

Equivalent to (9) is also

$$P(A') = \Pr\{|\, x - a\,| \geqslant X\} \leqslant \frac{s^2}{X^2}. \tag{10}$$

An alternative formulation of Tchebycheff's inequality is the following: if A denotes a range of x-values for which $|x - a| < ts$, where t is a given constant, a the mean value, and s^2 the variance, then

$$\Pr\{|x - a| < ts\} \geqq 1 - \frac{1}{t^2} \tag{9'}$$

$$\Pr\{|x - a| \geqq ts\} \leqq \frac{1}{t^2}. \tag{10'}$$

In the case of a distribution including finite jumps the same derivation holds and leads to Eq. (9), etc. But, it should be noted that if a jump of magnitude α occurs at the point $a + X$, the value of $P(A)$ is given by

$$P(A) = P(a + X) - P(a - X) - \alpha,$$

since A does not include its end points according to our definition. The equality sign in (9) can hold only in the case of an arithmetical probability (see Problem 9).

The great importance of Tchebycheff's inequality lies in its universal validity. For this very reason it will rarely provide the sharpest results in individual cases. It serves as an important tool in proofs and general estimates.

Problem 1. An arithmetical distribution is given by $p(x) = 2^{-m}\binom{m}{x}$, where x takes the values 0, 1, 2, ..., m. Show that the sum of the $m + 1$ probability values is 1. Plot the d.f. for $m = 6$.

Problem 2. The Poisson distribution is given by $e^{-c}c^x/x!$ where c is a positive number and x takes the values $x = 0, 1, 2, 3, \ldots$. Show that the sum of the $p(x)$ is 1 and plot for $c = 2$ the steps of the c.d.f. up to the point where P exceeds 0.99. Compute the mean value and variance.

Problem 3. Show that the function $p(x) = \lambda e^{-\lambda x}$ where λ is a positive constant and x goes from zero to $+\infty$ determines a density distribution. Compute the d.f. and plot its graph for $\lambda = 1.5$.

Problem 4. A density distribution equal to zero for $x \leqq 0$ has for positive x-values the form $p(x) = Cx^n e^{-\lambda x}$ with $\lambda > 0$ and n a positive integer. Compute the d.f. and determine the value of the constant C. Show that the graph of the c.d.f. has one inflection point and find its coordinates.

Problem 5. Let a and b be real positive constants and $b > a$. Assume that the density $p(x)$ is constant in the interval $(-a, +a)$ and drops linearly to zero in the intervals (a, b) and $(-a, -b)$. Find the value of the

constant density and determine mean value and variance of the distribution.

Problem 6. How would the answer to Problem 5 change if on both sides of the interval $(-a, a)$ the density $p(x)$ drops to zero exponentially rather than linearly, i.e., according to a law const $\cdot e^{\pm\lambda x}$?

Problem 7. Compute mean value and variance of the Poisson distribution of Problem 2.

Problem 8. Compute mean value and variance of the distribution given in Problem 4.

Problem 9. Apply Tchebycheff's inequality to a symmetrical three-point distribution and show that in this case the lower bound for $P(A)$ is actually reached.

Problem 10. Given a uniform probability density distribution spread out between $-b$ and b, what is the lower bound as determined by Tchebycheff's inequality for $\Pr\{-b \leqslant x \leqslant b\}$? Answer the same question for a symmetrical triangular distribution.

2. Expectation Relative to a Distribution. Stieltjes Integral

2.1. *Expectation.* Let $f(x_i)$ be a function, defined for all labels x_i of an arithmetical distribution $p(x_i)$, $i = 1, 2, \ldots$; in the continuous case we will in general require that $f(x)$ be defined and continuous in the domain where $p(x) > 0$.

The expression

$$E[f] = \sum_i f(x_i)p(x_i) \tag{11}$$

or

$$E[f] = \int f(x)p(x)\,dx \tag{12}$$

will be called the *expectation of f relative to the distribution* under consideration, provided the respective series or integral converges absolutely. Note that, for a given distribution, E depends on the function f; it is a so-called functional. The $f(x)$ in these formulas can be considered a random variable. Using the above definition, the mean value and variance of a distribution can be written as

$$a = E[x], \qquad s^2 = E[(x - a)^2] = E[x^2] - \{E[x]\}^2. \tag{13}$$

Hence, the mean value is the expectation of x, and the variance is the expectation of $(x - a)^2$.

In the arithmetic case $E[f]$ is the limit, as n tends to infinity, of the average value of the $f(x_i)$ obtained from the first n observations within a collective. Assume, for example, that the collective is a game of chance with the possible results (label values) $x_1, x_2, \ldots, x_k$ and let $f(x_i)$ be the reward due to a player whenever x_i comes up. Let S be the stake in each single turn. After n turns, the player's total reward is $\sum_{i=1}^{k} n_i f(x_i)$ and, therefore, his average balance $\sum_{i=1}^{k} (n_i/n) f(x_i) - S$, which in the limit gives $E[f] - S$. The equation $S = E[f]$ would, therefore, mean that in the long run the player neither loses nor wins. A game of chance in which $S = E[f]$ is customarily called *fair*. Conversely, the concept of expectation (at least relative to an arithmetical distribution) might be interpreted as the stake S in a fair game of chance. A similar interpretation is possible in the case of a continuous distribution if it is approximated in a suitable way by a sequence of arithmetic distributions.

The operator E is a *linear operator*; this means that

$$E[f_1(x) + f_2(x) + \cdots + f_r(x)] = E[f_1(x)] + E[f_2(x)] + \cdots + E[f_r(x)], \tag{14}$$

$$E[cf(x)] = cE[f(x)], \qquad \text{for } c = \text{constant}. \tag{15}$$

Relations (14) and (15) follow immediately from the definition of the symbol E in (11) and (12). Here, all expectations are with respect to the same one-dimensional distribution $p(x)$.

There is a simple relation between *the expectation* of $f(x)$ *relative to the distribution function* $P(x)$ *of* x *and the expectation of* $y = f(x)$ *relative to the distribution of* y. Denote by $\phi(y)$ the inverse function of $f(x)$ which we assume to be (piecewise) monotonic and differentiable. Regarding such a monotonic arc we can write the distribution of y in the form

$$G(y) = \Pr\{f(z) \leqslant y\} = \Pr\{z \leqslant \phi(y)\} = P(\phi(y)). \tag{16}$$

The expectation of y is then, in terms of a Stieltjes integral

$$\int y \, dG(y) = \int f(x) \, dP(x), \tag{17}$$

where the first integral is over the region of y which corresponds to the original region of x. This can be proved directly by means of the definitions (19) and (19′) which follow in the next subsection. We shall use (17), however, only in the cases covered by (17′) and (17″), which follow presently.

If $p(x)$ is the density of $P(x)$, our result is contained in the substitution formula

$$\int f(x)p(x)\,dx = \int yp[\phi(y)]\phi'(y)\,dy, \tag{17'}$$

where the integration region of y corresponds to that of x.

We consider now the case of an arithmetical distribution, which is of great interest. Consider first a finite number of attributes, e.g., $k = 10$, and probabilities $p_1, p_2, \ldots, p_{10}$ and a "gain" $f(x)$ connected with $x = 1, 2, \ldots, 10$, where, for example:

$$f(1) = f(3) = f(5), \qquad f(2) = f(4) = f(6), \qquad f(7) = f(9), \qquad f(8) = f(10).$$

Now set $f(1) = y_1, f(2) = y_2, f(7) = y_3, f(8) = y_4$; then we may write:

$$\begin{aligned} &f(1)p_1 + f(2)p_2 + \cdots + f(9)p_9 + f(10)p_{10} \\ &\quad = y_1(p_1 + p_3 + p_5) + y_2(p_2 + p_4 + p_6) + y_3(p_7 + p_9) + y_4(p_8 + p_{10}) \\ &\quad = y_1g_1 + y_2g_2 + y_3g_3 + y_4g_4\,. \end{aligned}$$

The distribution of the g_i, $i = 1, \ldots, 4$ is derived from that of the p_i by mixing.

Consider, as another example, a game with two dice, or, since "independence" is without importance, imagine urns filled with lots marked (1, 1), (1, 2), ..., (1, 6), (2, 1), ..., (2, 6), ..., (6, 6). We are interested in the expected value of the sum of the two results x and y; with $f(x, y) = x + y$ we have

$$\begin{aligned} E[(x+y)] = \sum_x \sum_y (x+y)p_{xy} = 2p_{11} + 3p_{12} + \cdots + 7p_{16} + 3p_{21} + \cdots \\ + 8p_{26} + \cdots + 12p_{66}\,. \end{aligned}$$

Obviously the same result is obtained in the form

$$2 \cdot p_{11} + 3(p_{12} + p_{21}) + 4(p_{13} + p_{22} + p_{31}) + \cdots + 12p_{66}\,.$$

Let us now consider an arithmetical distribution with countably many labels 1, 2, ... We assume that $E[f] = \sum_{i=1}^{\infty} f_i p_i$ exists and that

$$f(x_i) = f_1\,, \quad i \in A_1; \qquad f(x_i) = f_2\,, \quad i \in A_2\,, \cdots, \qquad f(x_i) = f_n\,, \quad i \in A_n\,, \cdots$$

and we form

$$E[f] = \sum_{\nu=1}^{\infty} f_\nu \sum_{i \in A_\nu} p_i = \sum_{i=1}^{\infty} f_i p_i\,. \tag{17''}$$

If the f are all positive, or more generally, if $\Sigma f_i p_i$ converges absolutely, the new form of $E[f]$ is a rearrangement of an absolutely convergent series. This case is in general all that is needed.

2.2. *Use of the Stieltjes integral.* The distribution function $P(x)$ enables us to describe both main types of distribution, the discrete or arithmetic one with finite jumps and the continuous or geometric one with continuous density, by the same formalism.[1] It can also be used to unify the notations (11) and (12).

Denote by $f(x)$ a *continuous* function[2] in $(a < x \leqslant b)$, where $f(a + 0)$ exists, and by $P(x)$ a d.f. Divide the interval $(a, b]$ into intervals $\delta_i : x_i < x \leqslant x_{i+1}$ by means of the points $a = x_0, x_1, x_2, \ldots, x_n, x_{n+1} = b$ and denote by L_i and l_i the upper and lower bounds[3] of $f(x)$ in the interval δ_i; put $L_i - l_i = \omega_i$ and form as in (38) of Chapter II

$$\bar{S} = \sum_{i=0}^{n} L_i \, \delta P_i \qquad \text{and} \qquad \underline{S} = \sum_{i=0}^{n} l_i \, \delta P_i , \tag{18}$$

where $\delta P_i = P(x_{i+1}) - P(x_i)$. Since $f(x)$ is uniformly continuous we can find η such that $\omega_i \leqslant \epsilon$ where $x_{i+1} - x_i < \eta$. Choosing n and the x_i such that $x_{i+1} - x_i < \eta$ for all i we have

$$\bar{S} - \underline{S} < \epsilon[P(b) - P(a)].$$

Let $n \to \infty$ and let at the same time the maximum length of δ_i tend to zero. Then $\bar{S}$ and $\underline{S}$ have a common limit which we denote by

$$\lim_{n\to\infty} \bar{S} = \lim_{n\to\infty} \underline{S} = \int_a^b f(x) \, dP(x). \tag{19}$$

If f is continuous in $(a, b]$ the integral to the right from a to b exists.

Likewise, if in every subinterval δ_i we take an arbitrary point ξ_i we again obtain, the limit being understood in the same sense as before,

$$\lim \sum_{i=0}^{n} f(\xi_i)[P(x_{i+1}) - P(x_i)] = \int_a^b f(x) \, dP(x), \tag{19'}$$

---

[1] For a useful textbook presentation, see W. RUDIN, *Principles of Mathematical Analysis*, Chapter 6. New York, 1953.

[2] The definition can be given for bounded $f(x)$ as in Section 7 of Chapter II. Here, however, we need only the case of continuous $f(x)$.

[3] The numbers L_i and l_i are not necessarily values of f if f is merely bounded; however, they are values of f if f is continuous.

since the sum on the left-hand side is included between $\underline{S}$ and $\bar{S}$. We call the right-hand side of (19) or of (19′) *the Riemann-Stieltjes integral of* $f(x)$ *over* $(a, b]$ *with respect to* $P(x)$.

We see that *if* $f(x)$ is continuous in $(a, b]$ the left-hand side of (19′) exists and is equal to the common limits of $\underline{S}$ and $\bar{S}$. From definitions (19) or (19′) the following properties follow, whenever the respective integrals exist:

(1) $\int_a^b f(x)\,dP(x) = -\int_b^a f(x)\,dP(x).$

(2) $\int_a^b f(x)\,dP(x) + \int_b^c f(x)\,dP(x) + \int_c^a f(x)\,dP(x) = 0.$

(3) $\int_a^b [f(x) + g(x)]\,dP(x) = \int_a^b f(x)\,dP(x) + \int_a^b g(x)\,dP(x).$

$$\int_a^b c f(x)\,dP(x) = c\int_a^b f(x)\,dP(x) \qquad (c \text{ a constant}).$$

(4) $\int_a^b f\,d(P_1 + P_2) = \int_a^b f(x)\,dP_1(x) + \int_a^b f(x)\,dP_2(x).$

(5) *If* $|f(x)| \leqslant M$, in $(a, b]$, *then*

$$\left|\int_a^b f\,dP\right| \leqslant M[P(b) - P(a)].$$

Consider now a sequence of d.f.'s $P_1(x)$, $P_2(x)$, ... which converges towards a d.f. $P(x)$ at every continuity point of $P(x)$. Let $f(x)$ be continuous in $(a, b]$ and a and b be continuity points of $P(x)$. Then

(6) $\lim_{n\to\infty} \int_a^b f(x)\,dP_n(x) = \int_a^b f(x)\,dP(x).$

This follows from definition (19).

Note the following comment. If the Lebesgue-Stieltjes integral $\int_a^b f\,dP$ exists the same holds for $|f|$ and

$$\left|\int_a^b f\,dP\right| \leqslant \int_a^b |f|\,dP.$$

This is not generally true for a Riemann Stieltjes integral.

In order always to have $\int_a^b dP = P(b) - P(a)$, we agree that the interval of integration is $a < x \leqslant b$. If then there are jumps at a and b we have

$$\int_a^b dP = \int_{a+0}^b dP = P(b) - P(a+0) = P(b) - P(a).$$

In Section 7.2 of Chapter II we considered the more general case of bounded $f(x)$ but under the assumption of continuous $P(x)$. If $f(x)$ is merely bounded the $\int_a^b f(x)\,dP(x)$ may exist even if $f(x)$ has a finite number of discontinuity points b_ν *provided* $P(x)$ *is continuous at each* b_ν.[4] We can then surround each b_ν by a subinterval i_ν which gives an arbitrarily small contribution to the sums (18).

Consider now two important particular cases.

(a) *If* $P(x)$ *has a continuous derivative,* $p(x) = P'(x)$ *everywhere,* or more generally if it is absolutely continuous, it follows by the mean value theorem that

$$P(x_{i+1}) - P(x_i) = P'(\xi_i)(x_{i+1} - x_i);$$

if then $f(x)$ is continuous we obtain from (19′) by means of the definition of the ordinary (Cauchy) integral for continuous f

$$\int_a^b f(x)\,dP(x) = \int_a^b f(x)P'(x)\,dx. \tag{20}$$

More generally, the Stieltjes integral exists if $P(x)$ is absolutely continuous and the right-hand side of (20) exists as a Riemann integral (see Chapter II, Section 7).

(b) If $P(x)$ is the d.f. of an arithmetical distribution which at points $a_1, a_2 \ldots$ in the interior of the interval (a, b) has steps $p(a_1), p(a_2), \ldots$ and no steps at a and b, each of the differences $p(x_{i+1}) - p(x_i)$ in (19′) vanishes unless this interval contains one of the label values a_k: if $x_i < a_k \leqslant x_{i+1}$, one has $P(x_{i+1}) - P(x_i) = p(a_k)$ and if $f(a_k) \neq 0$, the left-hand side of (19′) has the value $\sum_{k=1}^{\infty} f(a_k)\,p(a_k)$ so that

$$\int_a^b f(x)\,dP(x) = \sum_{k=1}^{\infty} f(a_k)p(a_k), \tag{21}$$

and this series converges.

For a $P(x)$ which combines the two particular cases, the Stieltjes integral reduces to the corresponding combination of ordinary integrals and series.

If the interval of integration is infinite we consider $\int_a^b f(x)\,dP(x)$ and simultaneously and independently we let $a \to -\infty$, $b \to +\infty$. If the limit

$$\lim_{a \to -\infty, b \to +\infty} \int_a^b f(x)\,dP(x)$$

[4] For the relation between this result and the one in Chapter II, Section 7.2, see Lebesgue [18], p. 276.

exists, this is the Riemann-Stieltjes integral of $f(x)$ with respect to $P(x)$ over the real axis; it is denoted by $\int_{-\infty}^{+\infty} f(x)\,dP(x) = \int f(x)\,dP(x)$. It can be proved that if $f(x)$ is continuous and bounded, this generalized integral exists.[5] Our formulas (11) and (12) take then the form

$$E[f] = \int f(x)\,dP(x); \tag{22}$$

this integral may also exist for unbounded $f(x)$, as in the case of moments, if for example $dP/dx = p(x)$ tends sufficiently strongly to zero as $|x| \to \infty$.

In the following, Stieltjes integrals with the element $dP(x)$ where $P(x)$ is a d.f. will be used in general simply as an abbreviated form of

$$\sum_x f(x)p(x) \qquad \text{or} \qquad \int f(x)p(x)\,dx,$$

or a combination of both. The sign $\int$ has the meaning $\int_{-\infty}^{+\infty}$. Similarly, $\int^x$ means integration over all label values smaller than x.

For some purposes, it is useful to introduce besides E [Eqs. (11), (12)] a second functional operator which can easily be derived from E. If f is a function of x, the expectation $E[f]$ is a constant and $(f - E[f])^2$ is again a function of x. The expectation of this latter function is then called the *variance of* f and written $\mathrm{Var}[f]$ or also $\mathrm{Var}(f)$:

$$\mathrm{Var}[f] = E[[f - E[f]]^2] = \int f(x)^2\,dP(x) - \left(\int f(x)\,dP(x)\right)^2. \tag{23}$$

For $f(x) = x$, $\mathrm{Var}[x]$ becomes identical with what was previously called the variance of the distribution $P(x)$. In the same way, we have called $E[x]$ the mean value of the distribution $P(x)$, and also the expectation of x. Note that $\mathrm{Var}[x]$ is not a linear operator.

In analogy to (7) and (7′) we have then (see Problem 11)

$$E[(f - f_0)^2] = \mathrm{Var}[f] + (E[f] - f_0)^2. \tag{24}$$

2.3. *Schwarz' inequality.* Consider

$$\int [\lambda f(x) + \mu g(x)]^2\,dP(x) = E[(\lambda f + \mu g)^2].$$

[5] Cf. a paper by M. FRÉCHET, "Sur quelques définitions possibles de l'intégrale de Stieltjes," *Duke Math. J.* 2 (1936), pp. 383–395.

We assume that $E[f(x)]$ and $E[g(x)]$ exist. Now the quadratic form in λ and μ:

$$E[(\lambda f + \mu g)^2] = \lambda^2 E[f^2] + 2\lambda\mu E[fg] + \mu^2 E[g^2]$$

is never negative; hence its discriminant is positive or zero:

$$(E[fg])^2 - E[f^2]E[g^2] \leqslant 0$$

or

$$(E[fg])^2 \leqslant E[f^2]E[g^2]. \tag{25}$$

This is Schwarz' inequality.

In the two particular cases of an arithmetic and of a geometric distribution we obtain

$$\left(\sum f_i g_i p_i\right)^2 \leqslant \left(\sum f_i^2 p_i\right) \cdot \left(\sum g_i^2 p_i\right)$$

and

$$\left(\int f(x)g(x)p(x)\,dx\right)^2 \leqslant \left(\int f^2(x)p(x)\,dx\right) \cdot \left(\int g^2(x)p(x)\,dx\right).$$

3. Generalizations of Tchebycheff's Inequality

3.1. *Immediate generalizations.* The Tchebycheff inequality (9) can be generalized in a rather obvious way, with respect to the general concept of variance (23). We derive the result by first proving the so-called *Markov*[1] *lemma*, which is of independent interest and may be formulated as follows.

Let $g(x)$ be a non-negative function of the chance variable x, $E[g]$ the expectation of $g(x)$ relative to a given distribution, and A the set of x-values for which $g(x) < t^2E[g]$, where t is a given constant. Then we have

$$P(A) \geqslant 1 - \frac{1}{t^2}. \tag{26}$$

Proof. The probability for the region where $g(x) \geqslant t^2E[g]$ is $P(A') = 1 - P(A)$. Computing $E[g]$ by (22) we obtain

$$E[g] = \int g(x)\,dP(x) \geqslant t^2E[g] \int_{(A')} dP = t^2E[g] \cdot P(A')$$

[1] This author's name appears as Markoff in his German publications. We have used the spelling Markov, the official American transliteration from the Russian, for consistency throughout this volume.

which gives

$$P(A') \leqslant \frac{1}{t^2} \tag{26'}$$

and this is equivalent to (26).

We obtain (9') and (10') for $g(x) = (x - a)^2$.

Denote now by $f(x)$ an arbitrary function of x and put $g(x) = [f - E[f]]^2$ then $E[g] = \mathrm{Var}[f]$ and the inequality $g(x) < t^2 E[g]$, which determines A, becomes $[f - E[f]]^2 < t^2 \mathrm{Var}[f]$, or

$$|f - E[f]| < t\,|\sqrt{\mathrm{Var}[f]}|, \tag{27}$$

For the region A determined by (27) *the inequality*

$$P(A) \geqslant 1 - \frac{1}{t^2}$$

holds which generalizes Tchebycheff's inequality. For the random variable $f(x)$, (27) *is* Tchebycheff's inequality.[2] The statement is empty for $t \leqslant 1$.

Another almost immediate generalization is the following: if $P(x)$ is a probability distribution, whose absolute moment of order m exists, a any real number and $m > 1$, then by the same reasoning which led us to (26), we have

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

This last integral is the probability for $|x - a| > X$ and we obtain

$$\Pr\{|x - a| \geqslant X\} \leqslant \frac{1}{X^m} \int |x - a|^m \, dP(x), \tag{10''}$$

which reduces to (10') for $m = 2$, a the mean value.

3.2. *Kolmogorov's inequality.* A wide generalization of Tchebycheff's inequality is due to Kolmogorov and is known as *Kolmogorov's inequality.*[3]

[2] We have seen (problem 9) that for a general distribution the Tchebycheff inequality cannot be improved. If we know, however, that the distribution is unimodal and continuous, a sharper estimate has already been given by Gauss (cf. Cramér [4], p. 256).

[3] This is really a statement on an n-dimensional distribution. We give it here, as a generalization of Tchebycheff's inequality, although n-dimensional distributions will be "officially" introduced only at the end of the chapter. Some readers may prefer to read Section 7 ff before studying the following derivation.

It will be used in the derivation of the so-called strong law of large numbers (Chapter V) as Tchebycheff's inequality is used in that of the (weak) law of large numbers (Chapters IV and V).

Let $x_1, x_2, \ldots, x_n$ *be independent random variables* (which means that the joint distribution $P(x_1, \ldots, x_n)$ of the n variables is a product of n one-dimensional distributions) *with mean values* a_k *and variances* r_k^2 *and put* $y_k = x_k - a_k$, $k = 1, 2 \ldots, n$; *let also*

$$x_1 + \cdots + x_k = X_k, \qquad a_1 + \cdots + a_k = b_k, \qquad r_1^2 + \cdots + r_k^2 = s_k^2, \tag{28}$$

and R *be a positive number. Then the probability* P *of the inequality*

$$| X_k - b_k |_{\max} \geqslant R, \qquad k = 1, 2, \cdots, n$$

is less than or equal to s_n^2/R^2, viz.,

$$P \equiv P(R, n) = \Pr\{| X_k - b_k |_{\max} \geqslant R, \quad k = 1, 2, \cdots, n\} \leqslant \frac{s_n^2}{R^2}. \tag{29}$$

Here $P(R, n)$ *is the probability that at least one of the* $| X_k - b_k |$ *is* $\geqslant R$. According to (29) all n deviations $| X_k - b_k |$ remain less than R with probability greater than or equal to $1 - (s_n^2/R^2)$.

For $n = 1$, this statement reduces to Tchebycheff's inequality.

Our proof is elementary. It is seen from (29) that we need a lower bound for $s_n^2 = E[(X_n - b_n)^2] = \int (X_n - b_n)^2 \, dP$, where this n-dimensional integral is over the whole n-dimensional label space of the $x_1, \ldots, x_n$. This space S can be divided into $(n + 1)$ non-overlapping and exhaustive n-dimensional regions $S_1, S_2, \ldots, S_n, S_{n+1}$, where S_k $(k = 1, 2, \ldots, n)$, is defined by the k inequalities:

$$\begin{aligned} S_k: &\quad | X_i - b_i | < R, \quad i = 1, 2, \cdots, k-1, \quad | X_k - b_k | \geqslant R \\ S_{n+1}: &\quad | X_i - b_i | < R, \quad i = 1, 2, \cdots, n. \end{aligned} \tag{30}$$

Then, with $Q_k = \int_{S_k} dP$, we have

$$Q_1 + Q_2 + \cdots + Q_n = P, \qquad Q_{n+1} = 1 - P. \tag{31}$$

Clearly,

$$s_n^2 = \int_{(S)} (X_n - b_n)^2 \, dP \geqslant \sum_{k=1}^{n} \int_{S_k} (X_n - b_n)^2 \, dP, \tag{32}$$

and we wish to estimate $\int_{S_k} (X_n - b_n)^2 \, dP$. Now, for $k \leqslant n$

$$(X_n - b_n)^2 = (X_k - b_k)^2 + 2 \sum_{j>k} (X_k - b_k) y_j + (y_{k+1} + \cdots + y_n)^2.$$

The variables y_j , $j > k$, do not occur in the definition of S_k , hence these variables are unrestricted in S_k and therefore, since $P(x_1, \ldots, x_n)$ is a product, we have for $j > k$, $i > k$

$$\int_{(S_k)} y_j \, dP = 0, \qquad \int_{(S_k)} X_k y_j \, dP = 0, \qquad \int_{(S_k)} y_i y_j \, dP = 0,$$

and hence

$$\int_{(S_k)} (X_n - b_n)^2 \, dP = \int_{(S_k)} (X_k - b_k)^2 \, dP + r_{k+1}^2 + \cdots + r_n^2, \qquad k = 1, \ldots, n. \tag{33}$$

But, by definition (30) of S_k :

$$\int_{(S_k)} (X_k - b_k)^2 \, dP \geqslant R^2 \int_{(S_k)} dP = R^2 Q_k ,$$

and, using this as well as (31), (32), and (33)

$$s_n^2 \geqslant \sum_{k=1}^{n} \int_{(S_k)} (X_n - b_n)^2 \, dP \geqslant \sum_{k=1}^{n} \int_{(S_k)} (X_k - b_k)^2 \, dP \geqslant R^2 \sum_{k=1}^{n} Q_k = R^2 P, \tag{34}$$

where P is the probability that at least one of the n values $| X_k - b_k |$ attains the value R; thus (29) has been proved.

4. Moments of a Distribution

We define the moment of order k about the point c, where k is an integer, as the expectation of the function $(x - c)^k$:

$$M_k^{(c)} = \int (x - c)^k \, dP(x) . \tag{35}$$

Obviously, the mean value a can be written: $a = M_1^{(0)}$, and the variance $s^2 = M_2^{(a)}$. Whenever, in the following, a moment is taken about the mean value, the superscript will be omitted:

$$M_k = \int (x - a)^k \, dP(x), \qquad M_2 = s^2. \tag{36}$$

Note the following relations:

$$M_0^{(c)} = 1 \qquad \text{for any } c; \tag{37}$$

$$M_1^{(c)} = a - c, \qquad \text{and in particular } M_1 = 0. \tag{38}$$

Equations (7) and (7′) may now be written

$$M_2^{(c)} = s^2 + (c - a)^2, \tag{39}$$

and

$$s^2 = M_2^{(0)} - a^2, \qquad M_2^{(0)} = s^2 + a^2. \tag{39′}$$

Equation (39) is the shift-of-origin rule [see Eq. (7)].

In a similar way, the moment $M_k^{(c)}$ may be expressed in terms of M_k, M_{k-1}, ... by applying the binomial theorem to $[(x - a) - (c - a)]^k$. In this manner one finds

$$M_k^{(c)} = M_k - \binom{k}{1}(c - a)M_{k-1} + \binom{k}{2}(c - a)^2 M_{k-2} - \cdots + (-1)^k(c - a)^k, \tag{40}$$

where $\binom{k}{\nu}$ are the usual binomial coefficients. If the arguments of $p(x)$ are 0, 1, 2, ... the *factorial moments* are defined as

$$M_{(k)} = \sum_x x(x - 1)(x - 2) \cdots (x - k + 1)p(x). \tag{41}$$

They are readily expressed in terms of moments about the origin.

For a *symmetrical distribution*, we have $dP(a + z) = -dP(a - z)$ for any z, and all moments of odd order vanish. The *moment of third order* may therefore be used to measure the degree of asymmetry of a distribution. In order to be independent of the scale used for the chance variable x, the "dimensionless" expression

$$r = \frac{1}{s^3} \int (x - a)^3 \, dP(x) \tag{42}$$

is employed for that purpose. The quantity r is called the *skewness* of the distribution.

To characterize a distribution still further, the moment of fourth order is sometimes used as follows. One introduces

$$K = \frac{1}{s^4} \int (x - a)^4 \, dP(x) - 3. \tag{43}$$

The quantity K is called the *kurtosis*; it has been designed so as to allow a comparison of the distribution under consideration with a certain standard distribution that is the subject of the following section. A more

theoretical justification for the choice of the "statistics" (42) and (43), will be found at the end of the next section in considerations which are of independent interest.

Much more on moments will be found in Chapter VIII.

Problem 11. Show that the shift of origin rule may be generalized as follows: $\int (f(x) - c)^2\, dP(x) = \mathrm{Var}[f] + (E[f] - c)^2$. In other words, using random-variable notation: $E[(y - c)^2] = \mathrm{Var}[y] + (E[y] - c)^2$ for any random variable y.

Problem 12. Compute the moment M_k of the distribution given by

$$p(x) = Cx^2e^{-x}, \quad x \geqslant 0; \qquad p(x) = 0, \quad x < 0.$$

Problem 13. Under what conditions can three real numbers be the moments $M_k^{(0)} = \int x^k\, dP(x)$ for $k = 0, 1, 2$ of a d.f. $P(x)$?

Problem 14. The absolute moment of order k about the origin of a distribution $P(x)$ is defined by

$$M_{|k|}^{(0)} = \int |x|^k\, dP(x).$$

What is the relation between absolute moments and ordinary moments in the case of a distribution symmetric with respect to the origin?

B. Gaussian Distribution, Poisson Distribution (Sections 5 and 6)

5. The Normal or Gaussian Distribution[1] in One Dimension

5.1. *Definitions.* The *normal distribution* or *Gaussian distribution* extends over all points of the x-axis; its density function is given by

$$p(x) = Ce^{-h^2(x-a)^2}, \qquad C > 0. \tag{44}$$

Since $\int p(x)\, dx = 1$, the two parameters C, h are connected by an equation. Introducing $u = h(x - a)$, one obtains

$$1 = \int p(x)\, dx = C \int e^{-u^2}\, d\left(\frac{u}{h} + a\right) = \frac{C}{h} \int e^{-u^2}\, du.$$

[1] We use the term "probability distribution" or more briefly "distribution" as a general term in both the discrete and the continuous cases.

The last integral is shown in calculus to have the value $\sqrt{\pi}$. Hence, $C = h/\sqrt{\pi}$ and (44) can be rewritten as

$$p(x) = \frac{h}{\sqrt{\pi}} e^{-h^2(x-a)^2}. \tag{45}$$

The maximum of the density function $p(x)$ lies at $x = a$. The line $x = a$ is the axis of symmetry of the bell-shaped density curve, i.e., $p(a - z) = p(a + z)$; obviously, $p(x) > 0$ for any finite x and $\lim_{x\to+\infty} p(x) = \lim_{x\to-\infty} p(x) = 0$. The magnitude of the parameter h determines the sharpness of the maximum at $x = a$, or the degree of concentration [see Fig. 3, where $p(x)$ is shown for three different values of h]. Sometimes, h is called the *measure of precision* of the normal distribution. Denoting by the letter ϕ the normal distribution with $h^2 = \frac{1}{2}$ and $a = 0$, we have

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \int \phi(x)\, dx = 1, \tag{46}$$

and

$$\int^0 \phi(x)\, dx = \int_0 \phi(x)\, dx = \tfrac{1}{2}.$$

We determine now the mean value and variance of the Gaussian distribution: using again

$$u = h(x - a), \qquad x = \frac{u}{h} + a, \qquad dx = \frac{1}{h} du, \tag{47}$$

the mean value is found as

$$\int xp(x)\, dx = a \int p(x)\, dx + \frac{1}{h} \int up(x)\, dx = a \cdot 1 + \frac{1}{h} \cdot \frac{h}{\sqrt{\pi}} \cdot \frac{1}{h} \int ue^{-u^2}\, du = a$$

since the second integral vanishes because the integrand is an odd function. This result has been anticipated by the use of the letter a in Eq. (44).

The variance is found through integration by parts:

$$\begin{aligned} s^2 = \int (x - a)^2 p(x)\, dx &= \frac{1}{h^2} \frac{h}{\sqrt{\pi}} \frac{1}{h} \int u^2 e^{-u^2}\, du \\ &= \frac{1}{h^2 \sqrt{\pi}} \frac{1}{2} \int ue^{-u^2}\, d(u^2) \\ &= \frac{1}{2h^2 \sqrt{\pi}} \left\{ \left[-ue^{-u^2} \right]_{-\infty}^{\infty} + \int e^{-u^2}\, du \right\} = \frac{1}{2h^2}. \end{aligned} \tag{48}$$

From now on, we use the variance s^2 rather than h^2 as the parameter and write

$$p(x) = \frac{1}{\sqrt{2\pi}s} e^{-(x-a)^2/2s^2} \tag{45'}$$

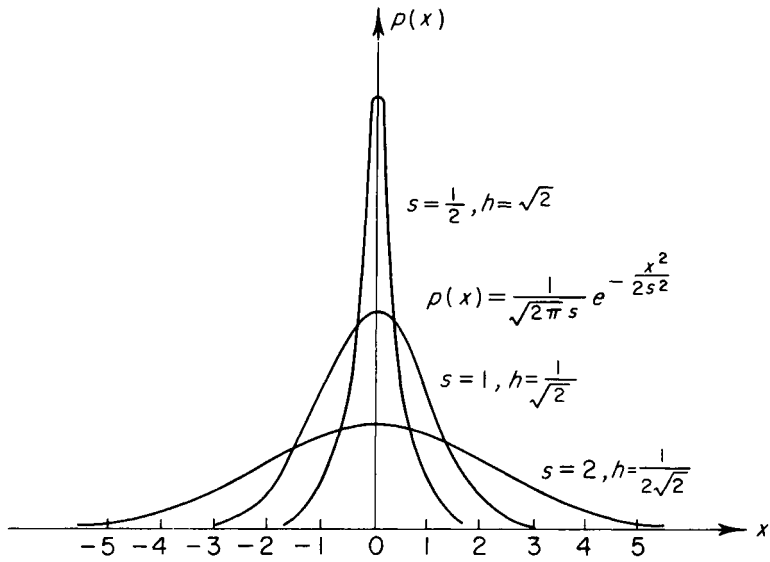

FIG. 3. Normal law. Densities.

(see Fig. 3); we see that $\phi(x)$ is obtained from $p(x)$ by taking $a = 0$, $s = 1$, and that

$$p(x) = \frac{1}{s}\phi\left(\frac{x-a}{s}\right). \tag{49}$$

The d.f. corresponding to $\phi(x)$ will be denoted by $\Phi(x)$:

$$\Phi(x) = \int^x \phi(z)\,dz = \frac{1}{\sqrt{2\pi}} \int^x e^{-z^2/2}\,dz\,. \tag{50}$$

The distribution function for the general Gaussian distribution $p(x)$ is

$$P(x) = \int^x p(z)\,dz = \frac{1}{\sqrt{2\pi}s} \int^x \exp\left(-\frac{(z-a)^2}{2s^2}\right) dz. \tag{51}$$

Writing now $(z - a)/s = t$, we find

$$P(x) = \frac{1}{\sqrt{2\pi}s}\int^{(x-a)/s} e^{-t^2/2}\, s\, dt = \frac{1}{\sqrt{2\pi}}\int^{(x-a)/s} e^{-t^2/2}\, dt.$$

Thus (see Fig. 4),

$$P(x) = \Phi\left(\frac{x - a}{s}\right). \tag{52}$$

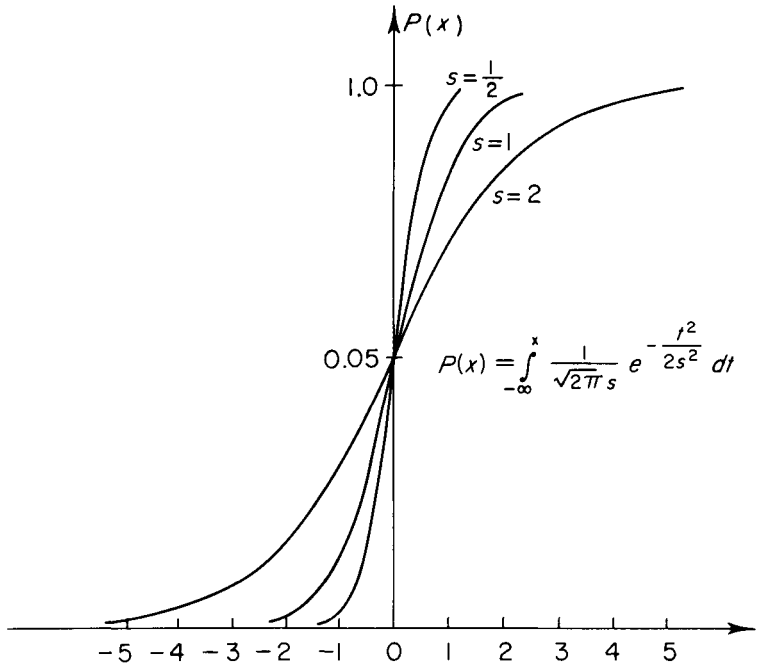

FIG. 4. Normal law. Cumulative distribution functions.

$\Phi(x)$ is a transcendental function and cannot be expressed in terms of more elementary functions. Tables of the Gaussian have been computed to a high degree of accuracy. Instead of Φ tables often given the function

$$G(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-z^2/2}\, dz = \Phi(x) - \tfrac{1}{2}. \tag{53}$$

See our Table I. Table II gives the inverse function $\Psi(x)$ to $\Phi(x)$.

In order to find an asymptotic expression for $G(x)$ for large x-values, consider the fraction

$$\frac{1 - 2G(x)}{(2/x)\phi(x)}.$$

For $x \to \infty$, numerator as well as denominator tend toward zero, hence l'Hospital's rule can be applied

$$\lim_{x\to\infty} \frac{1 - 2G(x)}{\frac{2}{x}\phi(x)} = \lim_{x\to\infty} \frac{-2\phi(x)}{2[-\frac{1}{x^2}\phi(x) - \phi(x)]}$$

$$= \lim_{x\to\infty} [1 + \frac{1}{x^2}]^{-1} = 1.$$

Thus, for large positive values of x, $G(x)$, and $\Phi(x)$ have the asymptotic approximations[2]

$$G(x) \sim \frac{1}{2} - \frac{\phi(x)}{x}; \qquad \phi(x) \sim 1 - \frac{\phi(x)}{x}. \tag{54}$$

For example, for $x = 4$, the respective expressions do not differ in the first five decimal places. It can be shown that the difference between the left- and the right-hand sides of (54) is of the form $z_1\phi(x)$ where $0 < z_1 < x^{-3}$. By partial integration of this last formula we obtain

$$\phi(x) = 1 - \frac{\phi(x)}{x} + \frac{\phi(x)}{x^3} + z_2\phi(x) \qquad (0 < z_2 < 3x^{-5}). \tag{54'}$$

It is sometimes of interest in statistics to specify a neighborhood of the mean value for a given symmetrical distribution, such that the probability for the chance variable x to fall in this neighborhood is $\frac{1}{2}$ (that is, such that the probabilities of belonging and of not belonging to the interval are equal). In the case of a normal distribution, the width of this "50-percent-chance neighborhood" which coincides with the interquartile range is, on account of symmetry, determined by

$$G(x) = \tfrac{1}{4}. \tag{55}$$

The solution of this equation is $x = 0.67449$. The half width of the 50-percent-chance neighborhood for the general normal distribution $p(x)$ is then:

$$z = 0.674s. \tag{56}$$

The values $a - 0.674s$, $a + 0.674s$ are sometimes called the *probable limits of a variable subject to the distribution* $N(a, s^2)$, i.e., to the normal distribution with mean a and variance s^2.

5.2. *Expansion of an arbitrary distribution.* We come now to the considerations mentioned at the end of the last section which, in addition

[2] The sign $\sim$ means that the limit of the quotient of the terms to the left and right tends toward 1 as $x \to \infty$; it also implies that for large x this quotient is close to unity.

to their independent interest, motivate definitions like skewness and kurtosis (42) and (43). One introduces the expansion of an "arbitrary function" into an infinite series whose terms are successive derivatives of the Gaussian distribution. This series is called *Charlier series*[3] or Gram-Charlier series or Edgeworth series in the English literature, *Bruns' series*[4] in the German literature.

Let $P(x)$ be a probability distribution, $\Phi(x)$ the Gaussian distribution (50) where $\Phi(-\infty) = 0$ and $\Phi(+\infty) = 1$. Then, $f(x) = P(x) - \Phi(x)$ represents the deviation of $P(x)$ from the Gaussian distribution $\Phi(x)$ and we propose to expand $f(x)$ into the series:

$$f(x) = P(x) - \Phi(x) = c_0\phi(x) + c_1\phi'(x) + c_2\phi''(x) + \cdots, \tag{57}$$

where $\phi(x) = \Phi'(x) = (1/\sqrt{2\pi})\, e^{-x^2/2}$, and $\phi'(x)$, $\phi''(x)$, ... are the successive derivatives of $\phi(x)$ (Fig. 5). By means of successive differentiation we obtain

$$\phi^{(n)}(x) = (-1)^n\phi(x)H_n(x), \tag{58}$$

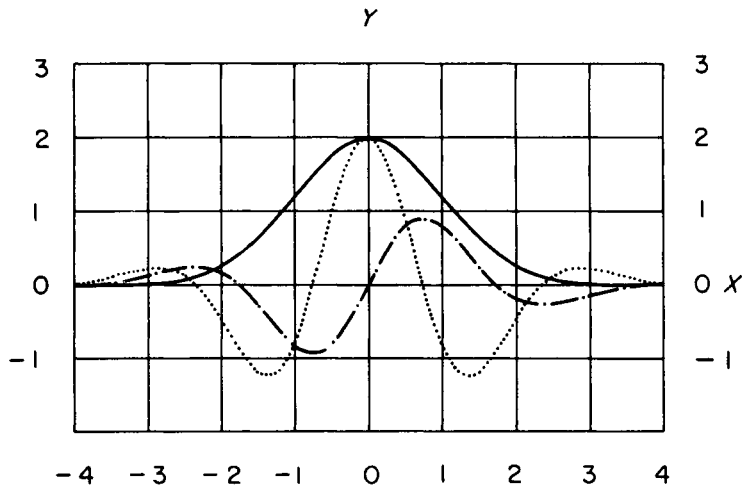

FIG. 5. Derivatives of normal distribution —— $= 5\phi$, —·— $= \frac{5\phi'''}{3}$, · · · $= \frac{5\phi^{(4)}}{3}$.

[3] C. V. L. CHARLIER, "Researches into the Theory of Probability," *Kungl. Fysiogr. Sällsk. i Lund. Handl.* **16** (1906).

F. Y. EDGEWORTH, "The Law of Error," *PCPS*, **20** (1905), p. 36.

[4] H. Bruns, *Wahrscheinlichkeitsrechnung und Kollektivmasslehre.* p. 39 ff. Leipzig, Berlin, 1906.

where for the H_n

$$H_{n+1}(x) = xH_n(x) - H_n'(x), \qquad \text{also} \qquad H_n' = nH_{n-1}\,. \tag{58'}$$

We obtain

$$\begin{aligned} &H_0(x) = 1, \quad H_1(x) = x, \quad H_2(x) = x^2 - 1, \quad H_3(x) = x^3 - 3x, \\ &H_4(x) = x^4 - 6x^2 + 3, \quad H_5(x) = x^5 - 10x^3 + 15x, \\ &H_6(x) = x^6 - 15x^4 + 45x^2 - 15, \ldots\,. \end{aligned} \tag{58''}$$

The $H_n(x)$ are the *Hermite polynomials*, or Hermite-Tchebycheff polynomials, which satisfy the orthogonality relations (to be verified by repeated integration by parts):

$$\int H_m(x)H_n(x)\phi(x)\,dx = \begin{cases} 0, & m \neq n \\ n! & m = n. \end{cases} \tag{59}$$

If (58) and (59) are used, the coefficients c_n in (57) can then be computed in the usual (Euler-Fourier) way and we find

$$c_n = \frac{(-1)^n}{n!}\int [P(x) - \Phi(x)]H_n(x)\,dx. \tag{60}$$

We have to assume that $P(x) - \Phi(x)$ tends to zero as $|\,x\,| \to \infty$ more strongly than the polynomials $H_n(x)$ tend to infinity. This is certainly true if $P(x)$ is the d.f. of an arithmetical distribution with a finite number of jumps since in this case, for $|\,x\,|$ greater than some finite value X, $P - \Phi$ equals $-\Phi$ or $1 - \Phi$ and these functions tend strongly toward the x-axis and the line $y = 1$, respectively.

If $P(x)$ is an arithmetic distribution[5] the formulas (60) reduce to finite sums. The reader may prove that

$$c_n = \frac{(-1)^{n+1}}{(n+1)!}\sum_{i=1}^{k} p_i H_{n+1}(a_i) \tag{61}$$

where jumps of magnitude p_i are at the points with abscissas a_i, $i = 1, 2, \ldots, k$.

If the arithmetical distribution has mean value zero and variance one, we obtain from (61) and (58''), $c_0 = 0$, $c_1 = 0$, $c_2 = -\frac{1}{6}M_3$, $c_3 = (1/4!)(M_4 - 3)$, ... where M_n are the moments of order n with respect to the mean value of the arithmetic distribution.

[5] For details of the computation and numerical examples, see v. Mises [21], pp. 250–265; in particular p. 255 ff.

In the case of a geometric distribution with density $p(x)$ we write

$$p(x) = \phi(x) + a_1\phi'(x) + a_2\phi''(x) + \cdots . \tag{62}$$

We assume $p(x)$ normed in such a way that the mean value is zero and the variance one. We assume that the Dirichlet conditions or other sufficient conditions of convergence hold in every finite interval and that $p(x)$ vanishes at infinity so strongly that $e^{x^2/2}p(x)$ is there of the order of a positive power[6] of $1/x$. We then obtain in the same way as before $a_n = (-1)^n \int p(x)\, H_n(x)\, dx$. Thus if M_n denote the moments about the origin of $p(x)$,

$$\begin{aligned} &a_0 = 1, \qquad a_1 = a_2 = 0, \qquad a_3 = -\frac{1}{6}M_3, \qquad a_4 = \frac{1}{24}(M_4 - 3), \\ &a_5 = \frac{1}{5!}(-M_5 + 10M_3), \qquad a_6 = \frac{1}{6!}(M_6 - 15M_4 + 30), \ldots \end{aligned} \tag{63}$$

and therefore

$$p(x) = \phi(x) - \frac{r}{6}\phi^{(3)}(x) + \frac{K}{24}\phi^{(4)}(x) + \cdots , \tag{62'}$$

where r and K are the skewness and kurtosis of $p(x)$ as in (42) and (43). By formal integration of (62) we obtain

$$P(x) - \Phi(x) = a_1\Phi'(x) + a_2\Phi''(x) + \cdots$$

and see comparing with (57) that the $a_n = c_{n-1}$ [as may also be seen by partial integration of (60)]. From (62') we then have

$$P(x) - \Phi(x) = -\frac{r}{6}\Phi^{(3)}(x) + \frac{K}{24}\Phi^{(4)}(x) + \cdots . \tag{62''}$$

Apart from the problem of convergence, for which sufficient conditions are known, the question of the practical value of these expansions as a useful approximation is controversial.[7]

---

[6] See Chapter IX by G. Szegoe in Frank-Mises, *Die Differential- und Integralgleichungnen der Mechanik und Physik*, Vol. 1, 2nd ed., p. 433 ff. Braunschweig, 1930. See also footnote 5, p. 136; and references in Cramér [4], p. 223 ff.

[7] See v. Mises [21], p. 264, regarding this question and also regarding sufficient convergence conditions. In this respect see also papers by v. Mises, 1912; G. Szegoe, 1926; W. Rotach, 1925; and others, quoted in Frank-Mises (*loc. cit.*), p. 433. See the discussion in Kendall [15], Vol. 1, p. 145 ff. See a paper by H. Cramér, "On some classes of series used in mathematical statistics," *Sixth Scandinavian Congress of Mathematics*, Copenhagen, 1925. The series (57) or (62) are also called Gram-Charlier series of Type A.

An asymptotic expansion different from (57) has been given by Edgeworth in 1905. It can be considered as a rearrangement of (57) and vice versa. It has desirable asymptotic properties. For its study see Cramér [5], Chapter VII, and the quotation in our Chapter IX, Section 6.1.

Problem 15. If the mean value of a normally distributed chance variable x is 4.5 and its standard deviation is 1.5 find the

(a) probability of x exceeding 6.0,
(b) probability of x falling into the range 3.0 to 6.0,
(c) the maximum value of the density,
(d) the density value at $x = 6.0$,
(e) the probable limits of x.

Problem 16. Prove that the following relation holds for the first three moments of a Gaussian distribution, irrespective of the point c about which the moment is taken

$$M_3^{(c)} - 3M_1^{(c)}M_2^{(c)} + 2(M_1^{(c)})^3 = 0.$$

Generalize!

Problem 17. Prove that for the Gaussian distribution, the moment of order $2m$ about the mean value equals

$$1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2m-1)\, s^{2m} \tag{a}$$

[this explains definition (43) whereby $K = 0$ for the normal distribution]. Show also that the absolute moment of odd order n taken about the mean value equals

$$\frac{(\sqrt{2}s)^n}{\sqrt{\pi}} \left(\frac{n-1}{2}\right)! \tag{b}$$

while that of even order is as in (a).

Problem 18. Show that if ξ is normally distributed with mean value zero and variance 1, $y = \xi^2$ has the density

$$g(y) = \frac{1}{\sqrt{2\pi y}}\, e^{-y/2},$$

(corresponding to $n = -\frac{1}{2}$, $\lambda = \frac{1}{2}$ in Problem 4).

6. The Poisson Distribution

We introduce the arithmetical probability distribution

$$\psi(x) = \frac{a^x}{x!}\, e^{-a}, \qquad a > 0, \qquad x = 0, 1, 2, \ldots . \tag{64}$$

One sees immediately that $\sum_{x=0}^{\infty} \psi(x) = 1$, since the infinite series with general term $a^n/n!$ is the power-series expansion of e^a. Next multiplying the equation $e^a = \sum_0^\infty a^x/x!$ by a, we find

$$ae^a = \sum_{x=0}^{\infty} \frac{a^{x+1}}{x!} = \sum_{x=1}^{\infty} x \frac{a^x}{x!} = \sum_{x=0}^{\infty} x \frac{a^x}{x!},$$

or

$$\sum_{x=0}^{\infty} x \frac{a^x}{x!} e^{-a} = \sum_{x=0}^{\infty} x\psi(x) = a. \tag{65}$$

Similarly,

$$a^2e^a = \sum_{x=0}^{\infty} \frac{a^{x+2}}{x!} = \sum_{x=2}^{\infty} x(x-1) \frac{a^x}{x!} = \sum_{x=0}^{\infty} x(x-1) \frac{a^x}{x!}.$$

Dividing again by e^a we find

$$\begin{aligned} a^2 &= \sum x^2\psi(x) - \sum x\psi(x) = (s^2 + a^2) - a; \\ s^2 &= a. \end{aligned} \tag{66}$$

We find in a similar way that

$$\sum_{x=0}^{\infty} x(x-1)(x-2) \cdots (x-\nu+1)\psi(x) = a^\nu, \qquad \nu = 1, 2, \cdots. \tag{67}$$

The expression to the left in (67) is the *factorial moment of order* ν of the distribution $\psi(x)$ [see Eq. (41)]. Clearly, the knowledge of the factorial moments of an arithmetic distribution up to a certain order is equivalent to that of the moments of the same order about any point.

Considering the quotient $\psi(x+1)/\psi(x)$, we find that the Poisson distribution first increases and then decreases: if a is not an integer, there is a single maximum value for $x = [a]$, where $[a]$ denotes the greatest integer less than a; if a is an integer, the values $\psi(a) = \psi(a-1)$ are the two equal maxima (see Table V).

An expansion somewhat analogous to (62) has been considered by Charlier: let $p(x)$ be an arithmetical distribution with jumps at $x = 0, 1, 2, \ldots$. We put in a formal way

$$p(x) = a_0\psi(x) + a_1\psi_1(x) + a_2\psi_2(x) + \cdots \tag{68}$$

where, with $\psi(x) = \psi_0(x)$, the $\psi_n(x)$ are defined as successive differences, rather than as derivatives:

$$\psi_n(x) = -\psi_{n-1}(x) + \psi_{n-1}(x-1), \qquad n = 1, 2, \cdots. \tag{69}$$

It is seen that, similarly to (58),

$$\psi_n(x) = \psi_0(x)p_n(x), \tag{70}$$

where the $p_n(x)$ are polynomials of degree n in x/a and where for the ψ_n (or the p_n) an orthogonality relation holds.[1]

Problem 19. Show that for the $p_n(x)$ defined in (70)

$$p_0(x) = 1, \qquad p_1(x) = \frac{x}{a} - 1, \qquad p_2(x) = \frac{x(x-1)}{a^2} - 2\frac{x}{a} + 1,$$

$$p_n(x) = \sum_{\nu=0}^{n} (-1)^{n-\nu} \binom{n}{\nu} \nu! \binom{x}{\nu} a^{-\nu}.$$

Problem 20. Prove the orthogonality relation

$$\sum_{x=0}^{\infty} p_n(x)\psi_m(x) = 0, \qquad m \neq n,$$

$$= \frac{n!}{a^n}, \qquad m = n.$$

C. Distributions in R_n (Sections 7 and 8)

7. Distributions in More Than One Dimension

7.1. *Arithmetic distributions.* To a k-dimensional collective there corresponds a k-dimensional distribution. Let us first discuss a simple example of a three-dimensional arithmetical distribution. Suppose three dice are thrown simultaneously, the label being the set (x, y, z) of the numbers that appear on the three upper faces; under usual conditions the three results independently assume the values $x = 1, 2, \ldots, 6$, $y = 1, 2, \ldots, 6$, $z = 1, 2, \ldots, 6$. There are altogether $6^3 = 216$ possible cases and they can be identified with the 216 lattice points inside and on the surface of the cube $x = 1, \ldots, 6$; $y = 1, \ldots, 6$; $z = 1, \ldots, 6$, in three-dimensional space. If the dice are correct, each point of the label space will carry the same probability $1/216$.

We may consider a space R_m of m dimensions with coordinates $x_1, x_2, \ldots, x_m$, instead of x, y, z. A point in this space is given as a real m-tuple $(x_1, x_2, \ldots, x_m)$. It is convenient to denote each m-tuple by a

[1] For more information regarding the expansion (68) and its convergence, see H. POLLACZEK-GEIRINGER, "Über die Poissonsche Verteilung und die Entwicklung willkürlicher Verteilungen," *Z. Angew. Math. Mech.* 8 (1928), pp. 292–309.

single symbol $\mathbf{x}$ (read: x-vector). Then we may say: each observation of an m-dimensional collective yields a certain vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ in m-dimensional space. The collective is discrete if there are altogether only enumerably many possible results $\mathbf{a}^{(1)}$, $\mathbf{a}^{(2)}$, ... occurring with the respective limiting frequencies $p_1, p_2, \ldots$ and

$$\sum_i p_i = \sum_i p(\mathbf{x}^{(i)}) = 1.^1 \tag{71}$$

In the case of a continuous m-dimensional collective, the probability density function $p(\mathbf{x}) = p(x_1, x_2, \ldots, x_m)$ is a function of m variables defined for the points of an m-dimensional continuum, and is continuous and non-negative everywhere except on certain surfaces of degree less than m, where finite jumps are admitted. For example $p(x, y) = \frac{1}{2}$ for $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 2$; $p(x, y) = 0$ everywhere else. More generally, a distribution in R_m may consist of m-dimensional densities $p(x_1, x_2, \ldots, x_m)$ defined in m-dimensional parts of R_m, of densities along surfaces of orders $(m-1)$, $(m-2)$, ..., 2, 1 in R_m and of concentrated probabilities p_i at countably many points. We have then to postulate that the m-dimensional Stieltjes integral over the whole infinite R_m equals unity.

7.2. *Distribution function.* Generalizing the concepts of a random variable, we may consider a random vector $\boldsymbol{\xi}$ with components $\xi_1, \xi_2, \ldots, \xi_m$ in R_m and explain its *probability distribution*: we denote by $P(A) = \Pr\{\boldsymbol{\xi} \in A\}$ the probability that $\boldsymbol{\xi}$ be contained in a region A of R_m; here $P(A)$ is an additive set function. In a manner similar to the one dimensional case, we introduce the d.f.

$$P(x_1, x_2, \cdots, x_m) = \Pr\{\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \ldots, \xi_m \leqslant x_m\}, \tag{72}$$

that is, the probability that $\xi_\mu \leqslant x_\mu$, $\mu = 1, 2, \ldots, m$. This $P(x_1, x_2, \ldots, x_m)$ is a function of the m real variables x_μ and possesses the following properties: *in each variable x_μ, P is a nondecreasing function, everywhere continuous, or at least continuous to the right, tending to zero as any $x_\mu \to -\infty$, and tending to one as all $x_\mu \to +\infty$.* These are necessary conditions for P to be a d.f. Again (see Chapter II, Section 7) there is a "converse" problem, the question whether a d.f. P which has the above

[1] The sum in (71) is written as a single sum. In the example of the 3 dice the i in (71) goes from 1 to 216 since there are altogether 216 possible results. If, in this example, the x, y, z would each take on countably many values (instead of six values) there would still be altogether only countably many possible results and (71) would hold.

properties uniquely determines a probability distribution in the space R_m such that $P(x_1, ..., x_m)$ represents the probability of $\xi_\mu \leqslant x_\mu$, $\mu = 1, 2, ..., m$. The answer is yes if probability is considered an abstract concept and is defined on the Borel sets of R_m. However, in a frequency theory of probability, the essentially singular part of $P(x_1, ..., x_m)$ is assumed equal to zero (see Chapter II, Section 7.2).

As before, we may consider a probability distribution to be analogous to a mass distribution, of total mass equal to one, distributed over the R_m. In the case of two variables x, y, the d.f. $P(x, y)$ denotes the mass to the left of the line $\xi = x$ and below the line $\eta = y$ (including the limits), and we have

$$\begin{aligned}\Pr\{a < \xi \leqslant b, \quad c < \eta \leqslant d\} &= [P(b, d) - P(b, c)] - [P(a, d) - P(a, c)] \\ &= P(b, d) - P(b, c) - P(a, d) + P(a, c) \geqslant 0,\end{aligned}$$

and similar formulas, as for example

$$\Pr\{a \leqslant \xi \leqslant b, \quad c < \eta \leqslant d\} = P(b, d) - P(b, c) - P(a-0, d) + P(a-0, c)$$

$$\begin{aligned}\Pr\{\xi > a, \quad \eta > c\} &= 1 - P(a, \infty) - P(\infty, c) + P(a, c) \\ &= 1 - P_1(a) - P_2(c) + P(a, c),\end{aligned}$$

where $P_1(a)$, $P_2(c)$ will be defined in Eq. (73). The conditions at infinity for P are for any x and y

$$P(x, -\infty) = P(-\infty, y) = 0, \qquad P(\infty, \infty) = 1.$$

We have used here the important concept of *marginal distributions* for the case $m = 2$. In fact, $P_1(x)$ is the probability that $\xi \leqslant x$, no matter what the value of η, and $P_2(y)$ is analogous.

$$P_1(x) = P(x, \infty), \qquad P_2(y) = P(\infty, y). \tag{73}$$

In the case of a distribution in R_m, we define in a similar way marginal distributions of orders 1, 2, ..., $(m - 1)$, respectively; there are $\binom{m}{\nu}$ marginal distributions of an order $\nu \leqslant m$. For example

$$P_{12...\nu}(x_1, x_2, ..., x_\nu) = P(x_1, x_2, ..., x_\nu, \infty, \infty, ..., \infty) \tag{74}$$

is the probability that $\xi_1 \leqslant x_1$, $\xi_2 \leqslant x_2$, ..., $\xi_\nu \leqslant x_\nu$ no matter what happens to the $(m - \nu)$ other random variables. The marginal distributions of order ν of a discrete or continuous distribution are, of course, distributions of the same type (discrete or continuous) in ν dimensions.

In the discrete case (in contrast to the procedure of p. 141, where

we enumerated the label points, one after the other) we may also write our formulas in terms of the m coordinates of the points $\mathbf{x}$. They are denoted by x_μ, $\mu = 1, 2, \ldots, m$, and each x_μ takes on certain labels: $x_\mu = a_1^{(\mu)}, a_2^{(\mu)}, \ldots$. Thus subscripts and superscripts are needed, and the sum of all probabilities which is equal to one, is written as an m-tuple sum. In the cases, $m = 2$, $m = 3$ we write in general $p(x, y)$ and $p(x, y, z)$ in order to avoid clumsy notation.

In the example of the three dice at the beginning of this section, each x_μ, $\mu = 1, 2, 3$ takes on the same values $1, 2, \ldots, 6$. The cumulative d.f. which had been defined in (72) is easy to find. For example, $P(4, 2, 5)$ is the probability of obtaining results where the first face shows 4 or less, the second 2 or less, the third 5 or less. If the 216 probabilities are equal, $P(4, 2, 5) = 40/216$. The marginal probability of the result "four with the first die, two with the second die" is

$$p_{12}(4, 2) = p(4, 2, 1) + p(4, 2, 2) + \cdots + p(4, 2, 6),$$

and $P_{12}(4, 2)$ the probability of at least 4 with first die and at least 2 with second die consists of 48 terms. Of course

$$P(6, 6, 6) = P_{12}(6, 6) = P_{23}(6, 6) = P_{13}(6, 6) = P_1(6) = P_2(6) = P_3(6) = 1.$$

In the continuous (or rather absolutely continuous) case the differentiable d.f. is related to the probability density by

$$P(x_1, x_2, \ldots, x_m) = \int^{x_1} \int^{x_2} \cdots \int^{x_m} p(t_1, t_2, \ldots, t_m)\, dt_1\, dt_2 \cdots dt_m . \tag{75}$$

Marginal distributions are now obtained by means of integrations; for example, for $m = 4$, we have

$$p_{13}(x_1, x_3) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2, x_3, x_4)\, dx_2\, dx_4 .$$

The probability that a result falls into an m-dimensional cell C, given by

$$c_1 \leqslant x_1 \leqslant d_1 , \qquad c_2 \leqslant x_2 \leqslant d_2 , \ldots, c_m \leqslant x_m \leqslant d_m ,$$

is

$$P(C) = \int_{c_1}^{d_1} \cdots \int_{c_m}^{d_m} p(x_1, x_2, \ldots, x_m)\, dx_1\, dx_2 \cdots dx_m . \tag{76}$$

An analogous formula holds for any region R of the label space.

7.3. *Conditional probabilities.* In Chapter I, Section 7 we introduced the concept of a *conditional distribution.* We considered (the notation here differs slightly from that in Chapter I) $P_A(B) = P(AB)/P(A)$, also denoted by $P(B \mid A)$, the probability of the set B if we already know that the result is in the set A, both A and B being subsets of the label space, AB being the intersection of A and B, and $p(A) > 0$. This is general. In the case of an arithmetical distribution $p(x, y)$ we had $p(y \mid x) = p(x, y)/p_1(x)$, the probability that the second label is y if we know that the first label is x. (In Chapter I we illustrated this by the tossing of two dice.)

Let us consider the continuous case. Let $p(x, y)$ be the probability density belonging to $P(x, y)$. We ask for the probability that the first result be between x and $x + h$, and the second result $\leqslant y$. This is

$$P(x + h, y) - P(x, y) = \int_x^{x+h} dx \int^y p(x, y)\, dy.$$

The probability that the first label be between x and $x + h$ is $P_1(x + h) - P_1(x)$ where $P_1(x)$ is the first marginal distribution function and

$$P_1(x + h) - P_1(x) = \int_x^{x+h} dx \int p(x, y)\, dy\,.$$

Then, the probability that the second label be $\leqslant y$ if we know that the first label is between x and $x + h$ is given by

$$\frac{P(x + h, y) - P(x, y)}{P_1(x + h) - P_1(x)} = \frac{\int_x^{x+h} dx \int^y p(x, y)\, dy}{\int_x^{x+h} dx \int p(x, y)\, dy}\,.$$

This means, in terms of mechanics, the mass in the vertical strip between the lines $\xi = x$ and $\xi = x + h$ and below the line $\eta = y$. We divide numerator and denominator by h and let $h \to 0$, assuming that $p_1(x) > 0$. Under customary assumptions we obtain

$$\lim_{h \to 0} \frac{P(x + h, y) - P(x, y)}{P_1(x + h) - P_1(x)} = \frac{(\partial P/\partial x)(x, y)}{p_1(x)} = \frac{\int^y p(x, y)\, dy}{p_1(x)}\,, \tag{77}$$

the mass along the line $\xi = x$ between $-\infty$ and y. This we denote as the *conditional probability that the second label be less than or equal to y if we know that the first label equals x*; it is also called the *conditional distribution function of y* if we know that the first label equals x.

If $p(x, y)$ is continuous in y, then $\int^y p(x, y)\, dy = (\partial P/\partial x)\,(x, y)$ may be differentiated with respect to y and we obtain *the conditional density of y* if we know that the first result is x:

$$p_2(y \mid x) = \frac{p(x, y)}{p_1(x)}, \tag{78}$$

a formula exactly like that applying to arithmetical probabilities. Similarly,

$$p_1(x \mid y) = \frac{p(x, y)}{p_2(y)}. \tag{78'}$$

We recover the case of independence, introduced in Chapter I, if $p_1(x \mid y) = p_1(x)$, then, from (78), $p(x, y) = p_1(x)p_2(y)$ and $p_2(y \mid x) = p_2(y)$. In terms of the two-dimensional c.d.f. $P(x, y)$ necessary and sufficient for independence is:

$$P(x, y) = P_1(x)P_2(y). \tag{78''}$$

The generalizations of these formulas and concepts in the case of more variables are left to the reader. More on independence is in Chapter XI.

7.4. *Various examples.*

(1) The random variables $x_1, \ldots, x_m$ are *uniformly distributed* in the parallelepiped $\mathbf{a} \leqslant \mathbf{x} \leqslant \mathbf{b}$ if the total "mass" is within this parallelepiped and the probability density is constant over all points of this volume. Then:

$$\begin{aligned} P(x_1, \ldots, x_m) &= 0 && \text{if} \quad x_i \leqslant a_i \quad \text{for any} \quad i \\ &= \prod_{i=1}^{m} \frac{y_i - a_i}{b_i - a_i} && \text{where} \quad y_i = x_i \quad \text{if all} \quad a_i \leqslant x_i \leqslant b_i \\ &= 1 && \text{where} \quad y_i = b_i \quad \text{if} \quad x_i > b_i . \end{aligned}$$

The corresponding density is given as

$$\begin{aligned} p(x_1, \ldots, x_m) &= 0 && \text{if} \quad \mathbf{x} \quad \text{outside the body} \\ &= \frac{1}{V} && \text{if} \quad \mathbf{x} \quad \text{inside this body of volume} \quad V = \prod_{i=1}^{m} (b_i - a_i). \end{aligned}$$

(2) The density of a two-dimensional Gaussian is given for $-\infty < \frac{x}{y} < \infty$, as:

$$p(x, y) = \frac{1}{2\pi s_1 s_2 \sqrt{1 - r^2}} \times \exp\left(-\frac{1}{2(1 - r^2)}\left[\frac{(x - a)^2}{s_1^2} - 2r\frac{(x - a)(y - b)}{s_1 s_2} + \frac{(y - b)^2}{s_2^2}\right]\right),$$

where a, b, s_1, s_2, r are five given constants. (See derivations and generalization in Chapter VIII.) It is immediately seen that $p(x, y)$ is constant on all points of the ellipses

$$\frac{(x-a)^2}{s_1^2} - 2r\frac{(x-a)(y-b)}{s_1 s_2} + \frac{(y-b)^2}{s_2^2} = k^2$$

(3) Denote by $f(z)$ a monotonic function, and by $P(x)$ a d.f. $P(x) = \Pr\{z \leqslant x\}$. We ask for the d.f. $G(u)$ of $f(z)$, i.e., $G(u) = \Pr\{f(z) \leqslant u\} = \Pr\{z \leqslant \phi(u)\}$ where $\phi(u)$ is the inverse function of $f(z)$.

Let, for example, $f(z) = az + b$; then

$$G(u) = \Pr\{az + b \leqslant u\} = \Pr\left\{z \leqslant \frac{u-b}{a}\right\} = P\left(\frac{u-b}{a}\right).$$

Or, let $f(z) = z^2$, which is piecewise monotonic. Then

$$\begin{aligned} G(u) = \Pr\{z^2 \leqslant u\} &= \Pr\{|z| \leqslant \sqrt{u}\} \\ &= P(\sqrt{u}) - P(-\sqrt{u}). \end{aligned}$$

(4) The analogous problem in the case of an m-dimensional distribution is analytically often very complicated.

Consider a density in two variables, $p(z_1, z_2)$. With $v_1 = v_1(z_1, z_2)$, $v_2 = v_2(z_1, z_2)$ we want to know the d.f. $G(u, v)$ of v_1, v_2. The formal solution is easily given

$$G(u, v) = \iint_{(R)} p(z_1, z_2)\, dz_1\, dz_2,$$

where the region R is defined by $v_1(z_1, z_2) \leqslant u$, $v_2(z_1, z_2) \leqslant v$. This, however, is often not easy to evaluate.

It is often more convenient to determine the probability density in the new variables. Assume that, in a certain region, the functions $v_1 = v_1(z_1, z_2)$, $v_2 = v_2(z_1, z_2)$ define a one-to-one correspondence between z_1, z_2 and v_1, v_2 such that the inverse functions $z_1 = z_1(v_1, v_2)$, $z_2 = z_2(v_1, v_2)$ exist; assume that the $v_i(z_1, z_2)$ have continuous derivatives $\partial v_i/\partial z_k$ at all points of the z_1, z_2-region (except perhaps at isolated points or along some isolated curves) such that the *Jacobian*

$$J = \frac{\partial(z_1, z_2)}{\partial(v_1, v_2)} = \begin{vmatrix} \dfrac{\partial z_1}{\partial v_1} & \dfrac{\partial z_2}{\partial v_1} \\ \dfrac{\partial z_1}{\partial v_2} & \dfrac{\partial z_2}{\partial v_2} \end{vmatrix}$$

is not infinite, its inverse $1/J = \partial(v_1, v_2)/\partial(z_1, z_2) \neq 0$. If A and B are regions in the z-plane and v-plane, such that these conditions hold we have (see texts on advanced calculus)

$$P(A) = \iint_{(A)} p(z_1, z_2)\, dz_1\, dz_2 = \iint_{(B)} p(z_1(v_1, v_2), z_2(v_1, v_2))|\, J\,|\, dv_1\, dv_2 .$$

In other words we have the probability element

$$p(z_1, z_2)\, dz_1\, dz_2 = p(z_1(v_1, v_2)\ z_2(v_1, v_2))\,|\, J\,|\, dv_1\, dv_2 .$$

[If to each (z_1, z_2) there corresponds one (v_1, v_2) but to a given (v_1, v_2) there is more than one point in the z-plane, we have to divide the region in the z-plane into several parts such that in each of them the correspondence is unique (see the second example (3), p. 146).]

Often we are interested only in the marginal probability $G(u)$, the d.f. of $v_1 = v_1(z_1, z_2)$. We may then put $v_2 = z_2$ and proceed as in the general case.

We illustrate by two examples:

(5) Let $p(x, y)$ be the two dimensional density of z_1, z_2. We ask for the d.f. of $z_1 + z_2$. This d.f. $G(u)$ is the probability that the point (z_1, z_2) lies in the half plane $z_1 + z_2 \leqslant u$, i.e.,

$$G(u) = \iint_{z_1+z_2\leqslant u} p(z_1, z_2)\, dz_1\, dz_2 = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{u-z_1} p(z_1, z_2)\, dz_2 \right) dz_1 .$$

With the new variable $z = z_1 + z_2$ we find

$$G(u) = \int_{-\infty}^{\infty} dz_1 \int_{-\infty}^{u} p(z_1, z - z_1)\, dz.$$

We shall find this problem in Chapter IV.

(6) We know the point density $p(x, y)$ of z_1, z_2. We wish to find the distribution $G(u)$ of the quotient z_1/z_2, i.e., the probability that $z_1/z_2 \leqslant u$, or in other words, the probability that the point z_1, z_2 lies in the region defined by $z_1/z_2 \leqslant u$. (The reader may sketch this region for $u = 2$, say. It consists of two parts of the plane adjacent at the origin.) By the considerations at the beginning of example (4) we obtain

$$G(u) = \iint_{z_1/z_2\leqslant u} p(z_1, z_2)\, dz_1\, dz_2$$

$$= \int_0^{\infty} dz_2 \left(\int_{-\infty}^{z_2 u} p(z_1, z_2)\, dz_1 \right) + \int_{-\infty}^{0} dz_2 \left(\int_{z_2 u}^{\infty} p(z_1, z_2)\, dz_1 \right) .$$

This gives for the density (we write z for z_2):

$$g(u) = \int_0^\infty zp(uz, z)\, dz - \int_{-\infty}^0 zp(uz, z)\, dz$$

and in the case of independence

$$g(u) = \int_0^\infty zp_1(uz)\, p_2(z)\, dz - \int_{-\infty}^0 zp_1(uz)\, p_2(z)\, dz.$$

The computation of the distribution of z_1/z_2 by the second method is in Chapter VIII, Section 7.2.

8. Mean Value and Variance in Several Dimensions

8.1. *Definitions.* We have to define *mean value* and *variance for an m-dimensional distribution.* For convenience, however, we shall write out most formulas for $m = 2$ only, and denote the variables by x and y. The mean value is an m-dimensional vector; for $m = 2$, in the notation used in (71), its components a and b are given by

$$a = \sum_i x^{(i)} p(x^{(i)}, y^{(i)}), \qquad b = \sum_i y^{(i)} p(x^{(i)}, y^{(i)}), \tag{79}$$

or, in the "two-dimensional" notation,

$$a = \sum_x \sum_y xp(x, y), \qquad b = \sum_x \sum_y yp(x, y) \tag{79'}$$

and for a geometrical distribution

$$a = \int\int xp(x, y)\, dx\, dy, \qquad b = \int\int yp(x, y)\, dx\, dy\,. \tag{80}$$

The point with coordinates (a, b) is the center of mass of a set of mass points or of a mass spread out continuously with a density $p(x, y)$. Also, any component of the mean value is the limit for $n \to \infty$ of the average of the values of that component obtained in the first n observations of our collective. Note that, for any m, the components of the mean value are identical with the respective mean values of the marginal distributions of first order; for example, for $m = 3$

$$a = \int\int\int xp(x, y, z)\, dx\, dy\, dz = \int x\, dx \int\int p(x, y, z)\, dy\, dz = \int xp_1(x)\, dx,$$

$$b = \int yp_2(y)\, dy, \qquad c = \int zp_3(z)\, dz. \tag{80'}$$

As to the variance it is seen that in the two-dimensional case, the following three expressions of second order can be formed:

$$s_{11} = \sum_x \sum_y (x - a)^2 p(x, y), \qquad s_{22} = \sum_x \sum_y (y - b)^2 p(x, y)$$

$$s_{12} = \sum_x \sum_y (x - a)(y - b) p(x, y) \tag{81}$$

or, in the continuous case,

$$s_{11} = \iint (x - a)^2 p(x, y)\, dx\, dy, \qquad s_{22} = \iint (y - b)^2 p(x, y)\, dx\, dy,$$

$$s_{12} = s_{21} = \iint (x - a)(y - b) p(x, y)\, dx\, dy. \tag{82}$$

Clearly, the shift-of-origin rules hold, namely, $s_{12} = \iint xy p(x, y)\, dx\, dy - ab$, etc. Again, as in (80′), $s_{11} = \int (x - a)^2 p_1(x)\, dx$, $s_{22} = \int (y - b)^2 p_2(y)\, dy$.

In the m-dimensional case, the mean value has m-components $a_1, a_2, \ldots, a_m$ and the quantities $s_{\mu\nu}$ can be arranged as a symmetrical square matrix $(s_{\mu,\nu})$; $\mu, \nu = 1, 2, \ldots, m$. The symmetry (i.e., the fact that $s_{\mu\nu} = s_{\nu\mu}$), follows directly from the definitions. Thus, it is seen that the extension of the notions of mean value and variance leads to the m-dimensional mean-value vector $\mathbf{a}$ or (a_μ), and to the m-by-m *matrix of the variances* $(s_{\mu,\nu})$, also called the *covariance matrix*. More specifically, we call the $s_{\nu\nu}$ *variances* and the $s_{\mu\nu}$, $\mu \neq \nu$, covariances.

The shift-of-origin rule generalizes to m dimensions in an obvious way:

$$s_{\mu\nu} = \int (x_\mu - a_\mu)(x_\nu - a_\nu)\, dP(\mathbf{x}) = \int x_\mu x_\nu\, dP(\mathbf{x}) - a_\mu a_\nu . \tag{83}$$

Finally, we prove that *the matrix of variances is non-negative definite.* Consider the expected value of $[\sum_{\mu=1}^m t_\mu(x_\mu - a_\mu)]^2$ which must be non-negative. Expanding, we find

$$E\left[\left[\sum_{\mu=1}^m t_\mu(x_\mu - a_\mu)\right]^2\right] = \sum_{\mu,\nu} s_{\mu\nu} t_\mu t_\nu . \tag{84}$$

We see that the second member is a non-negative quadratic form in the $t_1, t_2, \ldots, t_m$; this result is used in correlation theory. We call an m-dimensional distribution *singular* if the total mass lies in an hyperplane of less than m dimensions. If the distribution is nonsingular the matrix of variance is positive definite.

8.2. *Conditional mean value and conditional variance.*[1] Let $p(x, y)$ be a density. We may compute $E_2[g]$ the conditional expectation of a function $g(\xi, \eta)$ if we know that the first label equals x. This is the expectation of $g(x, \eta)$ with respect to the distribution $p(x, \eta)/p_1(x)$; hence $\int g(x, \eta)(p(x, \eta)/p_1(x)\, d\eta = E_2[g]$. In the same way, if we know that $\eta = y$, we denote by $E_1[g] = \int g(\xi, y)(p(\xi, y)/p_2(y)\, d\xi$ the expectation of $g(\xi, \eta)$ with respect to $p(\xi, y)/p_2(y)$.

We use in particular the *conditional mean* and the *conditional variance* (omitting the distinctions between x and ξ and between y and η). We define the expectation of x for given y which we denote as $\bar{x}(y)$:

$$\bar{x}(y) = E_1(x) = \frac{\int xp(x, y)\, dx}{p_2(y)}, \tag{85}$$

$$\bar{y}(x) = E_2[y] = \frac{\int yp(x, y)\, dy}{p_1(y)}. \tag{85'}$$

The right-hand side of (85) is a function of y; it reduces to a constant a in the case of independence

$$\frac{1}{p_2(y)} \int xp_1(x)p_2(y)\, dx = \int xp_1(x)\, dx = a = \iint xp(x, y)\, dx\, dy.$$

Likewise (85') reduces to the constant b in the case of independence. The notation E_1 means the expectation of x with respect to the distribution $p(x, y)/p_2(y)$, E_2 means the expectation with respect to the distribution $p(x, y)/p_1(x)$; and E will be, as always, reserved for the expectation with respect to $p(x, y)$. We shall also use the fact that $E[f(x)] = \iint f(x)\, p(x, y)\, dx\, dy = \int f(x)\, p_1(x)\, dx$ and likewise $E[g(y)] = \int g(y)\, p_2(y)\, dy$.

In the same way as (85) we introduce

$$\mathrm{Var}_1(x) = E_1[x^2] - \{E_1(x)\}^2. \tag{86}$$

Again, we understand by $\mathrm{Var}_1(x)$ the variance of the variable x with respect to the distribution $p(x, y)/p_2(y)$. Similarly, we have

$$\mathrm{Var}_2(y) = E_2[y^2] - \{E_2[y]\}^2 \tag{86'}$$

The analogous formulas in the case of an arithmetical distribution are obvious.

8.3. *Some equalities.* There are certain useful relations which we shall now derive, partly as a matter of exercise. We use the notation

[1] We shall return to these concepts in Chapter XI.

E, E_1, E_2 for expectations with respect to $p(x, y)$, to $p(x, y)/p_2(y)$ and to $p(x, y)/p_1(x)$.

(1) $E[E_1[f(x)]] = E[f(x)]$.

Here,

$$E_1[f(x)] = \frac{\int f(x)p(x, y)\,dx}{p_2(y)},$$

$$E[E_1[f(x)]] = \int\int \frac{f(x)p(x, y)\,dx}{p_2(y)} p_2(y)\,dy = E[f(x)].$$

Hence

$$E[E_1[f(x)]] = E[f(x)], \qquad E[E_2[g(y)]] = E[g(y)]. \tag{87}$$

(2) $$\begin{aligned} E_1[h(x)g(y)] &= g(y)E_1[h(x)] \\ E_2[h(x)g(y)] &= h(x)E_2[g(y)]. \end{aligned} \tag{88}$$

The left-hand side of the first Eq. (88) means

$$\int h(x)g(y)\frac{p(x, y)}{p_2(y)}\,dx = g(y)\int h(x)\frac{p(x, y)}{p_2(y)}\,dx.$$

(3) Let us compute:

$$E_1[h(x) + g(y)] = \int [h(x) + g(y)]\frac{p(x, y)}{p_2(y)}\,dx$$

$$= \int h(x)\frac{p(x, y)}{p_2(y)}\,dx + g(y)\int \frac{p(x, y)}{p_2(y)}\,dx = E_1[h(x)] + g(y).$$

Hence

$$\begin{aligned} E_1[h(x) + g(y)] &= E_1[h(x)] + g(y), \\ E_2[h(x) + g(y)] &= h(x) + E_2[g(y)]. \end{aligned} \tag{89}$$

(4) A more complicated formula follows. We wish to prove that (writing $\mathrm{Var}(y)$ rather than $\mathrm{Var}[y]$)

$$\mathrm{Var}(y) = E[\mathrm{Var}_2(y)] + \mathrm{Var}(E_2[y]). \tag{90}$$

Now:

$$\mathrm{Var}_2(y) = E_2[y^2] - [E_2[y]]^2,$$

$$E[\mathrm{Var}_2(y)] = E[E_2[y^2]] - E[(E_2[y])^2] = E[y^2] - E[(\bar{y}(x))^2] \tag{a}$$

$$\mathrm{Var}(E_2[y]) = \mathrm{Var}(\bar{y}(x)) = E[(\bar{y}(x))^2] - \{E[E_2[y]]\}^2$$

$$= E[(\bar{y}(x))^2] - (E[y])^2. \tag{b}$$

If we add the two last equations, (a) and (b), the second term to the right of (a) cancels against the first to the right of (b), and the statement follows.

8.4. *Remarks on expectations with respect to m-dimensional distributions.* The use of an m-dimensional Stieltjes integral affords a simplified formalism. Without giving here definitions of this integral for m dimensions we specify the meaning of the notation $E[f] = \int f(\mathbf{x})\, dP(\mathbf{x})$ in our two special cases,

$$E[f] = \int f(\mathbf{x})\, dP(\mathbf{x}) = \sum \cdots \sum f(x_1, x_2, \ldots, x_m) p(x_1, x_2, \ldots, x_m)$$

(arithmetical distribution)

$$\int\int \cdots \int f(x_1, x_2, \ldots, x_m) p(x_1, x_2, \ldots, x_m)\, dx_1\, dx_2 \cdots dx_m \tag{91}$$

(density).

If $\mathbf{v}$ denotes a vector with the components $v_1, v_2, \ldots, v_t$, $t \leqslant m$, and these components are functions of the label-space coordinates $x_1, x_2, \ldots, x_m$, the symbol $E[\mathbf{v}]$ is defined as a t-dimensional vector with the components

$$(E[\mathbf{v}])_\tau = \int v_\tau(\mathbf{x})\, dP(\mathbf{x}), \qquad \tau = 1, 2, \ldots, t. \tag{92}$$

We consider the case where $f(\mathbf{x})$ in (91) equals the sum $f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$ so that

$$\begin{aligned} &E[f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)] \\ &\quad = \int\int \cdots \int \{f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)\}\, dP(\mathbf{x}) \\ &\quad = \int\int \cdots \int f_1(x_1)\, dP(\mathbf{x}) + \cdots + \int\int \cdots \int f_m(x_m)\, dP(\mathbf{x}). \end{aligned}$$

This may be written as

$$E[f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)] = E[f_1(x_1)] + E[f_2(x_2)] + \cdots + E[f_m(x_m)]. \tag{93}$$

If in each of the m-dimensional integrals on the right-hand side we carry out $m - 1$ integrations and introduce the marginal distributions $P_i(x)$ we obtain with $\mathscr{E}_i[x] = \int x\, dP_i(x)$

$$E[f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)] = \mathscr{E}_1[f_1(x)] + \mathscr{E}_2[f_2(x)] + \cdots + \mathscr{E}_m[f_m(x)] \tag{94}$$

or in random variable notation

$$E[y_1 + y_2 + \cdots + y_m] = \mathscr{E}_1[y] + \mathscr{E}_2[y] + \cdots + \mathscr{E}_m[y]. \tag{94'}$$

Of course, these formulas express again the fact that E is a linear operator [see Eq. (14)].

We consider the expectation of a product of random variables assuming independence, viz.,

$$P(x_1, x_2, \ldots, x_m) = P_1(x_1)P_2(x_2) \cdots P_m(x_m). \tag{95}$$

(Note that if we know that the m-dimensional distribution resolves into the product of m one-dimensional distributions then these latter *must* be the marginal distributions of first order.) We have then, as seen immediately

$$E[f_1(x_1)f_2(x_2) \cdots f_m(x_m)] = \mathscr{E}_1[f_1[(x)]\, \mathscr{E}_2[f_2(x)] \cdots \mathscr{E}_m[f_m(x)]. \tag{96}$$

Tchebycheff's inequality can be extended to an m-dimensional distribution in various ways. Consider for example, the elements $s_{\mu\mu}$ in the main diagonal of the matrix of the variances. We have

$$\sum_{\mu=1}^{m} s_{\mu\mu} = \int \sum_{\substack{\mu=1 \\ |\mathbf{x}-\mathbf{a}|<r}}^{m} (x_\mu - a_\mu)^2\, dP(\mathbf{x}) + \int \sum_{\substack{\mu=1 \\ |\mathbf{x}-\mathbf{a}|\geqslant r}}^{m} (x_\mu - a_\mu)^2\, dP(\mathbf{x}). \tag{97}$$

The integrations here refer to the interior and exterior of a sphere with center a and arbitrary radius r. By omitting the first of the integrals in (97) and by replacing the sum of the squares in the integrand in the second integral by its minimum value r^2, we arrive at

$$\sum_{\mu=1}^{m} s_{\mu\mu} \geqslant r^2 \int_{|\mathbf{x}-\mathbf{a}|\geqslant r} dP(\mathbf{x}) = r^2(1 - P_r), \tag{98}$$

where P_r denotes $\Pr\{|\,\mathbf{x} - \mathbf{a}\,| < r\}$, that is, the probability to obtain a result which falls inside the sphere with center $\mathbf{a}$ and radius r. It follows that [compare with (9)]

$$P_r \geqslant 1 - \frac{\sum_{\mu=1}^{m} s_{\mu\mu}}{r^2}. \tag{99}$$

Thus, the probability that $(x_1 - a_1)^2 + (x_2 - a_2)^2 + \cdots + (x_m - a_m)^2 \leqslant r^2$ has a lower bound determined by r^2 and the sum $s_{11} + s_{22} + \cdots + s_{mm}^2$.

[2] See a generalization of this theorem in v. Mises [21], p. 62. A great many papers concern inequalities for the moments and various kinds of estimates. See also Chapter VIII. We mention also the monograph I. R. SAVAGE, "Probability inequalities of the Tchebycheff type," *Nat. Bureau of Standards Rept.* 1744 (1952) (replaced by a mimeographed one, March 1961, by Olkin and Savage). There is also a survey on Tchebycheff inequalities by H. J. GODURN, *J. Am. Statist. Assoc.* **50** (1955), pp. 932–945.

Additional facts on distributions in one and more dimensions will be discussed as they occur in special problems.

Problem 21. A two-dimensional chance variable is uniformly distributed (a) over the circle $x^2 + y^2 = R^2$; (*b*) over the rectangle $x = 0$, $x = c$, $y = 0$, $y = d$. Compute the mean values and the components of the variance.

Problem 22. Compute the constant K, the mean values, and the variance components for the distribution $p(x, y) = Ke^{-(cx+dy)}$, $c > 0$, $d > 0$; $x > 0$, $y > 0$.

Problem 23. The probability density for hitting a circular target is assumed to be proportional to $\exp(-c\sqrt{x^2 + y^2})$. Compute the probability for hitting an inner circle of radius R and compare this value with the result following from Tchebycheff's inequality.