CHAPTER V

SUMMATION OF CHANCE VARIABLES CHARACTERISTIC FUNCTION

A. Summation of Chance Variables and Laws of Large Numbers (Sections 1-4)

1. Summation of Chance Variables

1.1. *Statement of the problem.* Let us begin this chapter by restating the Bernoulli problem in a different form. Originally, we asked for the probability of drawing x ones, in n successive drawings, from an urn containing zeros and ones, the probabilities of drawing zero and one being p and q, respectively. Imagine now n urns, each of them containing zeros and ones, the corresponding probabilities p and q, where $p + q = 1$, being the same for each urn. We now draw once from each urn and ask for the probability that the sum of all results be x. Obviously, this is another way of formulating the Bernoulli problem, the difference being that the collectives K_1, K_2, ..., K_n (see Chapter IV, p. 167) are now directly given, instead of being derived by place selections from one given collective.

We can generalize the problem by assuming that in the n urns the probabilities are not the same, p_i, q_i being the values for the ith urn ($i = 1, 2, ..., n$). This case of n different alternatives is known as *Poisson's problem*. We shall, however, immediately admit still more general conditions, which include Poisson's problem as a special case.

Assume again n urns, but let each urn now represent a collective K_ν with a discrete distribution of a many-valued label. The possible label values can be any real numbers. In other words, a list of possible results of drawings from our urns would read as follows:

$$
\begin{array}{ll}
\text{possible results from } 1^{st} \text{ urn:} & A_1, A_2, ..., A_a \\
\text{possible results from } 2^{nd} \text{ urn:} & B_1, B_2, ..., B_b \\
\text{possible results from } 3^{rd} \text{ urn:} & C_1, C_2, ..., C_c \\
\dots\dots\dots\dots & \dots\dots
\end{array}
\qquad (1)
$$

The n collectives thus established are now combined into one collective with an n-dimensional distribution. The label in this collective is the possible result of n single drawings from the n urns, that is, an n-tuple of numbers $(x_1, x_2, ..., x_n)$, where x_ν is the result from the νth urn. Our problem is to determine the probability that the sum of the n drawings has the value x. This, evidently, requires a mixing operation within the n-dimensional collective.

Let us denote the given arithmetical distributions in the urns by $p_\nu(x)$, where $\nu = 1, 2, ..., n$. This means that, for instance, $p_1(x)$ vanishes for all values of x, except possibly for $x = A_1, A_2, ..., A_a$, and that $p_1(A_\alpha)$ is the probability of drawing A_α from the first urn, and

$$p_1(A_1) + p_1(A_2) + \cdots + p_1(A_a) = 1, \tag{2}$$

where the summation is to be extended over all labels of the first random variable. We shall each time use the symbol Σ_x for the summation over all x-values from $-\infty$ to ∞, being aware that *each such sum contains only a finite, or at most denumerable, number of non-vanishing terms.*

We may then write

$$\sum_x p_\nu(x) = 1, \qquad \nu = 1, 2, ..., n. \tag{3}$$

The probability of obtaining the sum x from n urns shall be denoted by $q_n(x)$:

$$q_n(x) = \Pr\{x_1 + x_2 + \cdots + x_n = x\}, \tag{4}$$

where x_ν is the result from the νth urn. Of course, the urns as such have no particular significance. We shall refer to $q_n(x)$ as the *probability that the sum of the chance variables in n independent collectives* (with discrete label sets) *is equal to x.* Since we perform n independent trials, $q_n(x)$ is the sum of all those n-tuple products $p_1(x_1)\, p_2(x_2) \cdots p_n(x_n)$ for which $x_1 + x_2 + \cdots + x_n = x$. Hence

$$q_n(x) = \sum_{x_1+\dots+x_n=x} \cdots \sum p_1(x_1)p_2(x_2) \cdots p_n(x_n). \tag{4'}$$

Later, we shall also admit non-discrete label sets.

The problem of the summation of chance variables is fundamental in probability theory. One example of an important application is the theory of observational errors (see also Chapter VII, Section 5), where one tries to determine the influence of errors, due to several independent causes, upon the accuracy of an observation. If, for example, somebody under-takes the precise determination of a supposedly constant voltage by means of repeated readings of a sensitive voltmeter, his readings will be subject to slight variations. These variations are, in general, due to the influence of several independent disturbances, such as the personal habits of the observer in reading his instrument, the influence of bearing friction, which is never quite constant, of slight changes in ambient temperature that will influence the sensitivity of the instrument, etc. We can assume that each of these causes determines the probability of a small, or "elementary" error. The actual observation is subject to the sum of all these errors.

1.2. *Solution of the problem.* Returning now to the mathematical problem of determining the function $q_n(x)$, we derive a recursion formula that connects q_n with q_{n-1}.

Let us call $q_{n-1}(y)$ the probability of drawing the sum y in $n-1$ single drawings from the first $n-1$ urns. If the first $n-1$ drawings have given the sum y, the total sum in n drawings will be x if the nth drawing yields $x-y$. The combined probability of obtaining y in the first $n-1$ drawings and $x-y$ in the nth is, under the assumption of independence, the product

$$q_{n-1}(y)p_n(x-y).$$

If this quantity is computed for each possible value of y and the results are summed, we obtain the probability of drawing the sum x in n drawings. (See also Chapter III, p. 147.) Hence, we arrive at

$$q_n(x) = \sum_y q_{n-1}(y)p_n(x-y). \tag{5}$$

Introducing $z = x - y$ as a new summation variable instead of y, we rewrite (5) as

$$q_n(x) = \sum_z q_{n-1}(x-z)p_n(z). \tag{6}$$

Often the term *convolution*[1] is used for this process of generating the sequence $\{q_n\}$ from $\{q_{n-1}\}$ and $\{p_n\}$. Equation (6) together with the obvious relation

$$q_1(x) = p_1(x) \tag{7}$$

[1] The corresponding term in German is *Faltung*.

solves our problem in principle, since it permits computation of q_2 from q_1, q_3 from q_2, and so on.

Let us give the definition of a convolution in a more general way. We may write $a(z)$, $b(z)$, $c(z)$ for $p_n(z)$, $q_{n-1}(z)$, $q_n(z)$. Then for positive arguments Eq. (6) reads

$$c(x) = a(0)b(x) + a(1)b(x-1) + \cdots + a(x)b(0). \tag{6'}$$

Even if the sequences $\{a_\kappa\}$ and $\{b_\kappa\}$ are not probability distributions we call the sequence $\{c_\kappa\}$ the convolution of $\{a_\kappa\}$, $\{b_\kappa\}$.

We give two examples of the operation of convolution. The first is immediate: if $p_\nu(0) = p$, $p_\nu(1) = q$, $\nu = 1, 2, \ldots, n$ are the probabilities of "failure" and "success" in one simple alternative, then the probability of obtaining x successes in n independent repetitions of this alternative is the probability that $x_1 + x_2 + \cdots + x_n = x$. Hence, the binomial distribution [Chapter IV, Eq. (11)] is the n-fold convolution of the simple alternative (see problem 1).

Next remember the "geometric" distribution (Chapter IV, p. 182). We saw that for simple Bernoulli trials qp^x is the probability that the first success happens at trial $x + 1$, or in other words qp^{x_1} is the probability of x_1 failures before the first success; similarly, we denote by x_2 the number of failures between the first and the second success, and generally by x_k the number of failures between the $(k-1)$th and kth success. Then the total number x of failures preceding the nth success equals $x = x_1 + x_2 + \cdots + x_n$ and the probability of this x is the probability of the sum $x_1 + x_2 + \cdots + x_n$ where each of these random variables has the geometric distribution. Thus, for positive integral n the negative binomial distribution $b_n(x) = \binom{-n}{x} q^n(-p)^x$, $x = 0, 1, \ldots$ is the probability of a sum generated by n-fold convolution from the geometric distribution.

Problem 1. Prove that the recursion formula (6) is satisfied in the Bernoulli problem, where

$$q_n(x) = \binom{n}{x} p^{n-x} q^x.$$

Problem 2. Use formula (6) to compute the probability $q_2(x)$ of getting the sum x in casting two dice each with the probabilities $p_1, p_2, \ldots, p_6$. Prove that in the case of correct dice

$$q_2(x) = q_2(14 - x) = \frac{x-1}{36} \quad \text{for} \quad x = 2, 3, \ldots, 7.$$

Problem 3. Find a solution analogous to that of Problem 2 for $q_3(x)$ in the case of 3 unbiased dice. Prove that

$$q_3(x) = q_3(21 - x) = \frac{\binom{x-1}{2}}{216}, \qquad x = 3, 4, ..., 8$$

$$q_3(9) = \frac{25}{216} = q_3(12)$$

$$q_3(10) = \frac{27}{216} = q_3(11).$$

Problem 4. If the values 0, 1, and 2 have the probabilities $p_\nu(0) = p$, $p_\nu(1) = q$, $p_\nu(2) = r$ for all ν, prove that the probability of obtaining the sum x in n trials is

$$q_n(x) = \sum_t \frac{n! p^{n-x+t} q^{x-2t} r^t}{t!(x-2t)!(n-x+t)!}.$$

2. The Laws of Large Numbers

2.1. *Mean value and variance of the distribution of a sum.* The first use which we make of the recursion formula (6) together with (7) is to compute mean value and variance of the distribution $q_n(x)$ in terms of the mean values and variances of the given distributions $p_\nu(x)$. Let us call a_ν and r_ν^2 mean value and variance[1] of $p_\nu(x)$:

$$a_\nu = \sum_x x p_\nu(x), \qquad r_\nu^2 = \sum_x (x - a_\nu)^2 p_\nu(x). \tag{8}$$

We call b_n the expectation of x relative to $q_n(x)$, that is, the expected mean value of the sum of n single drawings from n urns. Likewise, s_n^2 will be the variance of $q_n(x)$:

$$b_n = \sum_x x q_n(x), \qquad s_n^2 = \sum_x (x - b_n)^2 q_n(x). \tag{9}$$

To compute b_n we introduce (6) in the first expression (9):

$$b_n = \sum_x x \sum_z q_{n-1}(x - z) p_n(z) = \sum_{x,z} x q_{n-1}(x - z) p_n(z). \tag{10}$$

In this double sum, we replace the summation variable x by $y + z$ and

[1] If there are countably many labels in all or some of the $p_\nu(x)$ we assume that the respective a_ν and r_ν^2 exist.

(since $y = x - z$ takes every value if x and z do so independently) we obtain

$$b_n = \sum_{y,z} (y + z)q_{n-1}(y)p_n(z) = \sum_{y,z} yq_{n-1}(y)p_n(z) + \sum_{y,z} zq_{n-1}(y)p_n(z)$$

$$= \sum_y yq_{n-1}(y) \cdot \sum_z p_n(z) + \sum_y q_{n-1}(y) \cdot \sum_z zp_n(z). \tag{11}$$

But since $\Sigma_z p_n(z) = \Sigma_y q_{n-1}(y) = 1$ and the remaining sums are equal to b_{n-1} and a_n, we have found a recursion formula for b_n:

$$b_n = b_{n-1} + a_n . \tag{12}$$

Using this same formula to express b_{n-1} by b_{n-2}, etc., we arrive at $b_n = b_1 + a_2 + \cdots + a_n$. Since $b_1 = a_1$, as follows from (7), we have finally

$$b_n = a_1 + a_2 + \cdots + a_n , \tag{13}$$

or: *The mean value of a sum of independent chance variables equals the sum of the mean values of each single variable.*

Note that a_ν and a_μ refer in general to different distributions, if $\nu \neq \mu$.

Equation (13) is equivalent to

$$\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} (x_1 + x_2 + \cdots + x_n)p_1(x_1)p_2(x_2)\cdots p_n(x_n) = \sum_x xq_n(x), \tag{14}$$

where $q_n(x)$ is given by Eq. (6). Note that our result, derived here by means of Eq. (6) is a particular case of Eq. (94), Chapter III and is correct without assumption of independence. However, we restrict ourselves to independence in this section and in most of Chapters V and VI.

To determine now the variance $s_n{}^2$, we again use the recursion formula (6) and introduce $y = x - z$. If we also replace b_n by $b_{n-1} + a_n$, we have

$$s_n{}^2 = \sum_{x,z} (x - b_n)^2 q_{n-1}(x - z)p_n(z) = \sum_{y,z} [(y - b_{n-1}) + (z - a_n)]^2 q_{n-1}(y)p_n(z)$$

$$= \sum_y (y - b_{n-1})^2 q_{n-1}(y) \cdot \sum_z p_n(z) + \sum_z (z - a_n)^2 p_n(z) \cdot \sum_y q_{n-1}(y) \tag{15}$$

$$+ 2\sum_y (y - b_{n-1})q_{n-1}(y) \cdot \sum_z (z - a_n)p_n(z).$$

Here, the last term vanishes (both factors vanish); the remaining terms are s_{n-1}^2 and $r_n{}^2$. Thus

$$s_n{}^2 = s_{n-1}^2 + r_n{}^2, \tag{16}$$

and, consequently

$$s_n^2 = r_1^2 + r_2^2 + \cdots + r_n^2, \tag{17}$$

or: *The variance of a sum of independent chance variables equals the sum of the variances of each single variable.* In analogy to (14), we now also have

$$\sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} [(x_1 + x_2 + \cdots + x_n) - (a_1 + a_2 + \cdots + a_n)]^2 p_1(x_1) p_2(x_2) \cdots p_n(x_n)$$
$$= \sum_x (x - b_n)^2 q_n(x). \tag{18}$$

We reobtain previous results if we assume the n distributions $p_\nu(x)$ to be identical alternatives with label values 0 and 1. We have then $p_\nu(0) = p$, $p_\nu(1) = q$, $a_\nu = q$, $r_\nu^2 = pq$ for all ν's, and Eqs. (13) and (17) yield $b_n = nq$ and $s_n^2 = npq$.

2.2. *Law of large numbers for arithmetic distribution.* Equations (13) and (17) lead to what is called the *law of large numbers*, in Tchebycheff's form. This law refers to the distribution of $z = x/n$, that is, the arithmetical mean of the outcomes of n trials. If the probability of z is called $q_n'(z)$ it is clear that

$$q_n'(z) = q_n(nz) = q_n(x). \tag{19}$$

The expectation of z is

$$b_n' = E_n[z] = \sum_z z q_n'(z) = \frac{1}{n} \sum_x x q_n(x) = \frac{b_n}{n}, \tag{20}$$

and the variance of z

$$s_n'^2 = \mathrm{Var}[z] = \sum_z (z - b_n')^2 q_n'(z) = \frac{1}{n^2} \sum_x (x - b_n)^2 q_n(x) = \frac{s_n^2}{n^2}. \tag{21}$$

Applying Tchebycheff's inequality, we have

$$\Pr\{| z - b_n' | < \epsilon\} \geqslant 1 - \frac{s_n'^2}{\epsilon^2} = 1 - \frac{s_n^2}{n^2 \epsilon^2}. \tag{22}$$

Now, it was seen in (17) that s_n^2 equals the sum of the variances r_ν^2 of the originally given distributions $p_\nu(x)$. If all these r_ν^2 are bounded,

$$r^2 < r_\nu^2 < R^2, \tag{23}$$

their sum increases at the rate of n, that is, s_n^2/n remains a finite quantity and s_n^2/n^2 goes to zero with increasing n. The right-hand side of (21) then tends toward 1 however small ϵ is. In other words, the probability that z lies in any narrow vicinity of its expected value b_n' draws closer and closer to 1 as n increases. It is seen that the boundedness of r_ν^2 is a sufficient but not necessary condition for s_n^2/n^2 to approach zero. With b_n' defined by (20) we state the following *law of large numbers*:

The probability that the arithmetical mean of n independent chance variables falls in the interval $b_n' - \epsilon$, $b_n' + \epsilon$ *tends toward unity with increasing n, however small* ϵ *is, provided the variances of the individual chance variables are either bounded or increase in such a way that their arithmetical mean remains small as compared with n.*

This theorem has been proved so far only for the case of discrete probability distributions. It will be shown presently that it holds in all cases. Meanwhile, some other (less precise) forms of the statement may be mentioned. First, one is used to saying that an event is "almost certain" if its probability, considered as function of a parameter n, tends to 1. Thus, one could express the law of large numbers in these terms: For large n it is *almost certain* that the arithmetic mean of n observed values be found in the immediate vicinity of its expected value b_n'. Second, one may think of the probability distribution function of z, the function $Q_n'(z)$. The fact that its variance goes to zero means that the graph of the distribution function approaches more and more the graph of a unit-step function, with step at b_n'.

Another frequently used terminology is the following: Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be a sequence of random variables. One says (Cantelli, Fréchet) that ξ_n *converges in probability* to a constant c if for any $\epsilon > 0$ the probability of the relation $|\xi_n - c| > \epsilon$ tends to zero as $n \to \infty$. It is seen from (22) that $\xi_n = x/n - b_n'$ *converges in probability to zero.*

2.3. *Law of large numbers. General case.* Now let us take up the general case in which the individual chance variable is subject to a general probability distribution. The whole discussion can then be carried out by means of the cumulative distribution functions $P_\nu(x)$ instead of $p_\nu(x)$ and $Q_n(x)$ instead of $q_n(x)$.

Thus, instead of the $p_\nu(x)$, we consider the d.f's $P_\nu(x)$, $\nu = 1, 2, \ldots, n$ as given. $P_\nu(x)$ is defined as Pr{ label $\leqslant x$ in the νth collective}. Analogously, $Q_n(x) =$ Pr{ sum of n labels $\leqslant x$}. Thus, similarly to (4′)

$$Q_n(x) = \int_{x_1+\cdots+x_n\leqslant x} \cdots \int dP_1(x_1) \cdots dP_n(x_n). \tag{4''}$$

We wish again to give a recursion formula for $Q_n(x)$. If z is the result of the nth drawing, then the sum obtained in the first $n - 1$ drawings must be $\leqslant x - z$ in order to achieve a total sum not larger than x.[2] On the other hand, a sum $\leqslant x$ in n drawings cannot be achieved other than by combination of each possible result z in the last drawing with the result "not more than $x - z$" in the first $n - 1$ drawings. The probability $Q_n(x)$ will then be expressed by the Stieltjes integral

$$Q_n(x) = \int Q_{n-1}(x - z)\, dP_n(z). \tag{24}$$

This is the expression replacing the recursion formula (6) and it reduces to $Q_n(x) = \int Q_{n-1}(x - z)\, p_n(z)\, dz$ if $P_n(x)$ is a distribution with a probability density $p_n(x)$, to $Q_n(x) = \Sigma_z Q_{n-1}(x - z)\, p_n(z)$ if $P_n(x)$ is a step function whose steps are $p_n(x)$.

We now derive formulas (12) and (16) for general distributions $P_\nu(z)$. In the definition of the mean value $b_n = \int x\, dQ_n(x)$, the differential $dQ_n(x)$, the probability of obtaining in n drawings a sum between x and $x + dx$, is obtained when in (24) the factor $Q_{n-1}(x - z)$ is replaced by $dQ_{n-1}(x - z)$, where the differential refers to differentiation with respect to x at constant z. Thus the differential form of (24) is

$$dQ_n(x) = \int dQ_{n-1}(x - z)\, dP_n(z), \tag{24$'$}$$

and in the particular case of geometric distributions,

$$q_n(x) = \int q_{n-1}(x - z) p_n(z)\, dz. \tag{24$''$}$$

Writing $y + z$ for x, as in the derivation of (12) the mean value b_n becomes

$$\begin{aligned} b_n &= \int x \int dQ_{n-1}(x - z)\, dP_n(z) = \iint (y + z)\, dQ_{n-1}(y)\, dP_n(z) \\ &= \int y\, dQ_{n-1}(y) \cdot \int dP_n(z) + \int z\, dP_n(z) \cdot \int dQ_{n-1}(y) \\ &= b_{n-1} + a_n\,. \end{aligned} \tag{25}$$

Again, a relation similar to (14) holds; and again independence is not necessary. It follows also that, as in (16),

$$s_n^2 = \int (x - b_n)^2\, dQ_n(x) = s_{n-1}^2 + r_n^2\,.$$

[2] The summation letter z used here in the same way as in (10) and (11) has, of course, nothing to do with the arithmetical mean $z = x/n$, introduced on p. 228.

Thus it is seen that in the general case too, s_n^2 is the sum $r_1^2 + r_2^2 + \cdots + r_n^2$.

Since the transition from the sum x to the arithmetic mean z leads again to $s_n'^2 = s_n^2/n^2$, it is seen that the inequality (22) holds for the general case. In fact, in the law of large numbers as formulated on p. 231, no restriction was placed on the type of distribution to which the individual chance variables are subjected.

The law of large numbers is a first step in the analysis of the distribution $Q_n(x)$ or $Q_n'(z)$ for indefinitely increasing n. It is a statement of the increasing "concentration" of $Q_n'(z)$ at its mean value. In the next two sections we add some further investigations regarding the law of large numbers. In Chapter VI we shall develop an asymptotic expression for the distribution $Q_n(x)$ for $n \to \infty$. For that purpose it is necessary to make use of certain mathematical tools of a more advanced character, some of which will be considered in Part B of the present Chapter.

Problem 5. In a game of chance, the possible results are either a gain of \$100 or a loss of \$10, 20, 30, 40, each of these cases having the probability 0.20. Find by means of Tchebycheff's inequality a bound for the number of times this game must be repeated to have a chance of at least 95% that in the final outcome the average gain or loss per game will not exceed $\pm$\$20.

Problem 6. Compute expectation and variance of the sum of spots thrown with a correct die in n trials. If $n = 100$, what is the lower bound for the probability that the mean of the results lies between 3.4 and 3.6? Solve the same problem for $n = 1000$ and $n = 10{,}000$.

Problem 7. Consider two independent variables x_1, x_2 each having the Cauchy distribution:

$$P(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x, \qquad p(x) = \frac{1}{\pi(1 + x^2)}.$$

Show that $\frac{1}{2}(x_1 + x_2)$ has the same distribution; generalize to $(x_1 + x_2 + \cdots + x_n)/n$ where each of the independently distributed variables has the Cauchy distribution.

Problem 8. Compute $q_2(x)$ if $p_i(x) = [\Gamma(m_i + 1)]^{-1} x^{m_i} e^{-x}$; $i = 1, 2$; $x \geqslant 0$. Find also $q_n(x)$ for n distribution $p_i(x)$.

Problem 9. Marks are placed at random on the straight segment from $x = 0$ to $x = c$. The number of marks is n. For each single placing the probability density is constant: $p(x) = 1/c$. If z is the abscissa of the centroid of the n marks, what is the expectation and the variance of z?

Compute, for $n = 100$, the lower Tchebycheff bound for the probability that z lies between $E[z] - 0.05c$ and $E[z] + 0.05c$.

Problem 10. The n chance variables $x_1, x_2, ..., x_n$ are subject to probability distributions with the densities

$$p_\nu(x) = \frac{1}{\sqrt{\nu\pi}} \exp\left(-\frac{(x-\nu)^2}{\nu}\right), \qquad \nu = 1, 2, ..., n.$$

Compute the mean value and the variance of the distribution of the sum $x_1 + x_2 + \cdots + x_n$. Is the law of large numbers valid in this case?

Problem 11. Prove that in the summation of independent chance variables the central moments of third order M_3 are additive (like the mean values and the variances) but that this does not hold for M_4, M_5, etc.

3. Laws of Large Numbers Continued. Khintchine's Theorem. Markov's Theorem

Khintchine has proved that for n independent random variables which have one and the same distribution $P(x)$, *$z - b_n'$ converges in probability to zero if merely the mean value of the common distribution* exists,[1] without any assumption regarding its variance. See proof Sect. 8.5.

We return to the sum $x = x_1 + x_2 + \cdots + x_n$, *without assuming that the variables $x_1, ..., x_n$ are independent.* As said before, the derivation of Eq. (94), Chapter III, shows that for the mean value of x we obtain again as in (13) $b_n = a_1 + a_2 + \cdots + a_n$. It is clear that Eq. (17) does not hold for dependent variables since $\iint[(x_1-a_1)+(x_2-a_2)]^2 p_{12}(x_1, x_2)\, dx_1\, dx_2$ where $p_{12}(x_1, x_2)$ is a marginal distribution of order two of $p(x_1, x_2, ..., x_n)$ reduces to $\int (x_1 - a_1)^2 p_1(x_1)\, dx_1 + \int (x_2 - a_2)^2 p_2(x_2)\, dx_2 + 2\iint (x_1 - a_1)(x_2 - a_2)\, p_{12}(x_1, x_2)\, dx_1 dx_2 = r_1^2 + r_2^2 + c_{12}$ where this last term is called the covariance of x_1 and x_2. Denote by B_n the variance of the sum $x_1 + x_2 + \cdots + x_n$. Then, independent of the validity of (17) Tchebycheff's inequality yields

$$\Pr\{|z - b_n'| < \epsilon\} \geqslant 1 - \frac{B_n}{n^2\epsilon^2}.$$

Suppose now

$$B_n/n^2 \to 0, \tag{26}$$

[1] M. A. Khintchine, "Sur la loi des grands nombres." *C. R. Acad. Sci. (Paris)* **189** (1929), pp. 477–479. He gives two proofs, one using the method of truncation" due to Markov, the other that of characteristic functions.

as n increases indefinitely; then for an arbitrarily small η, a number n_0 can be found such that the inequality $B_n/n^2\epsilon^2 < \eta$ will hold for $n > n_0$ and consequently $P \geqslant 1 - \eta$ for $n > n_0$. *Hence, $z - b_n'$ converges to zero in probability if the number n of independent or dependent chance variables tend to infinity and* B_n/n^2 tends to zero as $n \to \infty$. This *more general law of large numbers*, which includes the theorem of p. 231, is due to Markov.

It is of interest that the sufficient condition $B_n/n^2 \to 0$ is also necessary if the n variables $x_i - a_i$, $i = 1, ..., n$ are uniformly bounded; that is, if a constant C, independent of n, can be found such that all $(x_i - a_i)^2 < C^2$. Assume for simplicity $a_i = 0$; then if P is the probability that $(x_1 + x_2 + \cdots + x_n)^2 < n^2\epsilon^2$, then $1 - P$ is the probability that $(x_1 + x_2 + \cdots + x_n)^2 \geqslant n^2\epsilon^2$. It follows that for $B_n = E[(x_1 + x_2 + \cdots + x_n)^2]$ the inequality

$$B_n < n^2C^2(1 - P) + n^2\epsilon^2P \tag{27}$$

holds; or

$$\frac{B_n}{n^2} < C^2(1 - P) + \epsilon^2P < C^2(1 - P) + \epsilon^2.$$

If Markov's theorem (that is, the law of large numbers) holds, $1 - P$ converges to 0 as $n \to \infty$; hence, the right-hand term can be made as small as we please and $B_n/n^2 \to 0$ follows indeed.

Consider independent variables. If for any of them the variance does not exist the quantity B_n has no meaning. However, the following important theorem holds, again due to Markov.[2] Assume without loss of generality that the mean values a_i are zero. *The law of large numbers holds if for any $\delta > 0$ the absolute moments of order* $1 + \delta$, i.e.,

$$E[|x_i|^{1+\delta}], \qquad i = 1, 2, \ldots .$$

all exist and are bounded.

Consider the following instructive example, given by Uspensky [26]: n variables have the same arithmetic probability distribution, with labels $a_m = 2^{m-2\log m}$ and probabilities 2^{-m} for $m = 1, 2, 3, \ldots$. The mean value of this distribution is

$$a = \sum a_m p(a_m) = \sum 2^{-2\log m} = 1 + \frac{1}{2^{\log 4}} + \frac{1}{2^{\log 9}} + \cdots .$$

The series is convergent, and therefore the expectation exists. Thus, from Khintchine's theorem, the law of large numbers holds. However, we cannot get this result from Markov's first theorem, for the variance of the distribution is

$$r^2 = \sum a_m{}^2 p(a_m) - a^2$$

[2] See a proof in Uspensky [26], p. 191 ff.

and the series

$$\sum a_m{}^2 p(a_m) = \sum 2^{m-4\log m}$$

is divergent; hence the variance r^2 does not exist and Markov's theorem does not apply. The reader may verify that the stronger theorem just given (without proof) does not apply either.

In Chapter XII we shall consider a broad generalization of the law of large numbers. We shall prove that the phenomenon of concentration, first described by Bernoulli, holds not only for the average of n results—if n is sufficiently large—but for an extensive class of functions of the n observations. At the present moment, however, we shall not consider this type of generalization. We turn now to the so-called strong law of large numbers—which, by the way, will be shown to hold equally in that general case.

4. Strong Laws of Large Numbers

4.1. The problem. The various forms of the law of large numbers considered in Chapter IV, Section 4 and in the last two sections of the present chapter were all of the following type: It is most probable that the frequency (or the average) computed from a long sequence of observations lies very close to the corresponding expected value (which is the probability of the single trial or a theoretical mean value). The following statement is stronger and more useful for applications: It is to be expected, with a probability approaching one, that in a long sequence of observations all deviations between observed and theoretical frequency lie, from a certain n on, below a given bound. More precise formulations will follow. In contrast to this strong law of large numbers, the law of large numbers is often called "weak" law of large numbers.

Theorems of this type were first given by Borel and by Cantelli[1] (see Section 4.4) for Bernoulli trials. A more general theorem was found independently by Pólya.[2] Here we follow the formulation and proof of Kolmogorov,[3] whose general theorem is based on his inequality which we proved in Chapter III, Section 3.2. For the convenience of the reader we restate here Kolmogorov's inequality.

[1] F. P. Cantelli, "Sulla probabilità come limite della frequenza." *Rend. Accad. Naz. dei. Lincei* [5] **26** (1917), p. 39 ff.

[2] G. Pólya, "Eine Ergänzung zu dem Bernoullischen Satz der Wahrscheinlichkeitsrechnung." *Nachr. Akad. Wiss. Göttingen. Math.-Phys. Kl.* (1921).

[3] A. Kolmogorov, "Sur la loi forte des grands nombres." *C. R. Acad. Sci. (Paris)* **191** (1930), p. 910 ff.

See a proof along similar lines in R. v. Mises, "Die Gesetze der grossen Zahl für statistische Funktionen." *Monatsh. Math. u. Physik* **43** (1936), pp. 105–178.

Let $P \equiv P(R, n)$ be the probability that for at least one ν between 1 *and n the deviation $| X_\nu - b_\nu |$ between the observed and the expected sums of results be $\geqslant R$. Then, with s_n^2 denoting the variance of the sum of n results*

$$P(R, n) \leqslant s_n^2/R^2.$$

To understand more easily the meaning of the theorem, as well as its proof, which is elementary once Kolmogorov's inequality is available, we consider first the case of n mutually independent alternatives: $p_i = p_i(0)$, $q_i = p_i(1)$.

4.2. *Strong law for alternatives.* There are n independent random variables which take on the values 0 or 1 with probabilities p_i, q_i. The number of "successes," or, the sum of the first ν results is $X_\nu = x_1 + x_2 + \cdots + x_\nu$ with expected value $\nu\bar{q}_\nu = q_1 + q_2 + \cdots + q_\nu$ and variance $s_\nu^2 = p_1q_1 + p_2q_2 + \cdots + p_\nu q_\nu$. Then with $X_\nu/\nu = z\nu$:

If $n + m$ alternatives with the success-probabilities $q_1, q_2, \ldots, q_{n+m}$ are observed, then the probability P, that for at least one subscript $\nu = n$, $n + 1, \ldots, n + m$ the *observed frequency z_ν of the first ν results deviates by ϵ or more from its expected value $\bar{q}_\nu$ is less than $1/n\epsilon^2$ independent of m*[4]:

$$P \leqslant \frac{1}{n\epsilon^2}. \tag{28}$$

This bound is only four times as high as that obtained from (22) viz., $s_n^2/n^2\epsilon^2 = (p_1q_1 + \cdots + p_nq_n)/n^2\epsilon^2 \leqslant n/4n^2\epsilon^2$, although the statement there refers to *one* subscript, $\nu = n$, only.

To prove the theorem it is sufficient to consider the case $m = n$ and to show that for $n + m = 2n$ the probability P_n, that for at least one ν between n and $2n$ the deviation $| X_\nu - \nu\bar{q}_\nu | \geqslant \nu\epsilon$, is less than $\frac{1}{2}n\epsilon^2$. In fact, an arbitrary section from n to $n + m$ can always be covered by a sequence of sections from n to $2n$, from $2n$ to $4n$, $4n$ to $8n$, etc., and the probability of a deviation $\geqslant\epsilon$ for at least one of these sections is certainly not greater than the sum of the probabilities for the single sections, viz., $P_n + P_{2n} + P_{4n} + \cdots$. This last sum, however, will have the bound

$$\frac{1}{2n\epsilon^2} + \frac{1}{4n\epsilon^2} + \frac{1}{8n\epsilon^2} + \cdots = \frac{1}{2n\epsilon^2}\left[1 + \frac{1}{2} + \frac{1}{4} + \cdots\right] \leqslant \frac{1}{n\epsilon^2}.$$

[4] Note that the notation differs from that in Section 2.

It therefore remains to estimate P_n. Now,[5] with $X_\nu = \nu z_\nu = x_1 + x_2 + \cdots + x_\nu$:

$$\begin{aligned} P_n &= \Pr\{\max | X_\nu - \nu\bar{q}_\nu | \geqslant \nu\epsilon; \quad \nu = n, n+1, \ldots, 2n\} \\ &\leqslant \Pr\{\max | X_\nu - \nu\bar{q}_\nu | \geqslant n\epsilon; \quad \nu = n, n+1, \ldots, 2n\} \\ &\leqslant \Pr\{\max | X_\nu - \nu\bar{q}_\nu | \geqslant n\epsilon; \quad \nu = 1, 2, \ldots, 2n\}. \end{aligned}$$

To this last probability we apply Kolmogorov's inequality, with the R and $s_n{}^2$ of Chapter IV replaced by $n\epsilon$ and $s_{2n}^2 = \Sigma_{\nu=1}^{2n} p_\nu q_\nu$, respectively, and obtain

$$P_n \leqslant \frac{1}{n^2\epsilon^2} \sum_{\nu=1}^{2n} p_\nu q_\nu \leqslant \frac{1}{n^2\epsilon^2} \cdot 2n \cdot \frac{1}{4} = \frac{1}{2n\epsilon^2}, \tag{29}$$

and the theorem is proved. It is clear that the proof and the estimate work in a similar way if the variances $r_k{}^2$ of the independent random variables are uniformly bounded $| r_k{}^2 | \leqslant r^2$. Then we obtain

$$P_n \leqslant \frac{1}{n^2\epsilon^2} \cdot 2nr^2 = \frac{2r^2}{n^2\epsilon^2}, \tag{29'}$$

and $P \leqslant 4r^2/n\epsilon^2$ results.

4.3. *General case.* Now we consider the general case with $b_\nu = E[X_\nu]$, $P_n = \Pr\{\max | X_\nu - b_\nu | \geqslant \nu\epsilon;\ \nu = n,\ n+1, \ldots, 2n\}$. We have, just as in (29)

$$P_n \leqslant \frac{1}{n^2\epsilon^2} s_{2n}^2 = \frac{4}{(2n)^2\epsilon^2} (r_1{}^2 + r_2{}^2 + \cdots + r_{2n}^2). \tag{30}$$

To show that for any given δ, an n, dependent on ϵ and δ, can be found so that for given k, $P_n + P_{2n} + P_{4n} + \cdots + P_{2^k n} < \delta$, it is necessary and sufficient to prove that the series $P_1 + P_2 + P_4 + \cdots$ converges. Now by (30) and since $1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \cdots = \frac{4}{3}$ we have

$$\begin{aligned} &P_1 + P_2 + P_4 + P_8 + \cdots \\ &\leqslant \frac{4}{\epsilon^2}\left[\frac{r_1{}^2 + r_2{}^2}{2^2} + \frac{r_1{}^2 + \cdots + r_4{}^2}{4^2} + \frac{r_1{}^2 + \cdots + r_8{}^2}{8^2} + \cdots\right] \\ &= \frac{4}{\epsilon^2}\left[\frac{r_1{}^2 + r_2{}^2}{4}\cdot\frac{4}{3} + \frac{r_3{}^2 + r_4{}^2}{16}\cdot\frac{4}{3} + \frac{r_5{}^2 + \cdots + r_8{}^2}{8^2}\cdot\frac{4}{3} + \cdots\right] \qquad (31) \\ &\leqslant \frac{16}{3\epsilon^2}\sum_{k=1}^{\infty}\frac{r_k{}^2}{k^2}. \end{aligned}$$

[5] The use of "max$| X_\nu - \nu\bar{q}_\nu |$" is not essential; it is rather a way to avoid saying each time the probability that "any of the deviations $| X_\nu - \nu\bar{q}_\nu |$ be greater than ... ."

Thus, we have obtained the *strong law of large numbers* as follows:

If $(n + m)$ *independent chance variables* x_i *with expectations* a_i *and variances* r_i^2, $i = 1, 2, \ldots, n$ *are observed and*

$$\sum_{k=1}^{\infty} \frac{r_k^2}{k^2} < \infty, \tag{32}$$

then, with $x_1 + \cdots + x_\nu = X_\nu$, $a_1 + \cdots + a_\nu = b_\nu$, *the probability* P *that for at least one* $\nu = n, n + 1, \ldots, n + m$, *the average deviation* $(1/\nu) \mid X_\nu - b_\nu \mid$ *be greater than or equal to* ϵ, *is less than an arbitrarily given* δ, *if* n (dependent on δ and ϵ but not on m) *is chosen large enough.*[6]

This is indeed a much stronger statement then *Tchebycheff's law of large numbers* which is obtained for $m = 0$.

It can be seen that the Markov condition, $B_n/n^2 \to 0$, which was sufficient for the weak law of large numbers, is satisfied if (32) holds.

The considerations regarding the statistical meaning of the law of large numbers presented in Chapter IV apply in the same way to the present stronger statement. Again, our statement refers to a great number of "experiments," each experiment consisting of $n + m$ single observations; however, our characterization of these experiments is now much more complete than before, since it relates to $n + m$ simultaneous inequalities valid for the results X_ν, stating that in the overwhelming majority of experiments none of those inequalities will hold. Sequences like 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, ... and others which do not satisfy Bernoulli's theorem (Chapter IV p. 175 ff) *a fortiori* do not satisfy the strong law of large numbers.

4.4. *Discussion.* We shall now investigate certain statements of the strong law which go beyond the formulations given here. For the following analysis it is sufficient to consider the simplest case (Section 4.2) and with all probabilities equal to $\frac{1}{2}$. Let $\epsilon > 0$, n and m be positive integers, and denote by $A^\epsilon_{n,m}$ the set of sequences for which at least once $\mid z_{n+\nu} - \frac{1}{2} \mid \geqslant \epsilon, \nu = 1, 2, \ldots, m$. Then, with the definitions of Chapter II, $A^\epsilon_{n,m}$ are sets of F, i.e., sets of the field of the collective of Chapter I, admitting probabilities $P(A^\epsilon_{n,m})$ and (28) is the probability statement

$$P(A^\epsilon_{n,m}) \leqslant \frac{1}{n\epsilon^2}\,. \tag{28}$$

[6] A. Khintchine studied the strong law of large numbers without the hypothesis of independence of the distributions. A. KHINTCHINE, "Sur la loi forte des grands nombres." *C. R. Acad. Sci.* (*Paris*) **186** (1928), pp. 285–287; "Remarques sur les suites d'événements obéissant à la loi des grands nombres." *Math. Sb.* **39** (1932), p. 115 ff.

Now, consider the sequence of sets $A^{\epsilon}_{n,m} \subset A^{\epsilon}_{n,m+1} \subset A^{\epsilon}_{n,m+2} \subset \ldots$ which is convergent. Denote the limit set by A^{ϵ}_{n}. Here, A^{ϵ}_{n} is the set of those sequences for which at least once $| z_k - \frac{1}{2} | \geqslant \epsilon$ for any $k > n$. The set A^{ϵ}_{n} is neither in F nor in F_1; it has measure but no content, and clearly, admits no verification. We have

$$\lim_{m\to\infty} P(A^{\epsilon}_{n,m}) = | A^{\epsilon}_{n} | \leqslant \frac{1}{n\epsilon^2}. \tag{28_1}$$

This is the first formally mathematical extension, where $| A^{\epsilon}_{n} |$ is a measure, but not a probability.

As in Chapter II, denote by F_2 the field of Lebesgue measurable sets and consider the sequence of sets in $F_2 - F_1$, $A^{\epsilon}_{n} \supset A^{\epsilon}_{n+1} \supset A^{\epsilon}_{n+2} \supset \cdots$, which likewise converges. We call the limit set A^{ϵ}. Here, A^{ϵ} is the set of all those sequences for which after any arbitrarily large number of trials, the inequality $| z_k - \frac{1}{2} | \geqslant \epsilon$ still holds, at least once; i.e., it is the set of sequences for which this inequality holds infinitely often. Since $| A^{\epsilon}_{n} | \leqslant 1/n\epsilon^2$ we can conclude that

$$\lim_{n\to\infty} | A^{\epsilon}_{n} | = | A^{\epsilon} | = 0, \tag{28_2}$$

or, equivalently: *The set of sequences for which the inequality under consideration holds only finitely often has measure one.*[7]

A last step is to consider a null-sequence $\epsilon_1 > \epsilon_2 > \epsilon_3 > \ldots$, and the corresponding sequence of sets $A^{\epsilon_1} \subset A^{\epsilon_2} \subset A^{\epsilon_3} \subset \cdots$, which also converges. Call A^0 the limit set. A^0 is the set of all sequences for which the relative frequency does not approach $\frac{1}{2}$. On account of $| A^{\epsilon_\nu} | \to 0$ we obtain

$$| A^0 | = \lim_{\nu\to\infty} | A^{\epsilon_\nu} | = 0. \tag{28_3}$$

The set of those sequences for which the observed relative frequencies approach $\frac{1}{2}$ has measure one.

If, as in this book, we consider probability as the limit of relative frequencies, then the law (28) is indeed a strong statement on actual occurrence, which goes beyond the (weak) law of large numbers and far beyond our "first postulate" of Chapter I, on the existence of limits of relative frequencies. If, on the other hand, a formal measure-theoretical definition of probability is adopted, then (28) is a purely analytic state-

[7] This result which is not a probability statement, in our sense, but a theorem of measure theory is often denoted as the strong law of large numbers.

ment (see discussion in Chapter IV, Section 4.2). The *statements* (28_1), (28_2), (28_3) *are in both instances statements of measure theory with no relation to real occurrence.*

It should thus be clear that any claim that, in the framework of a measure theory of probability, the "first postulate" can be proved or replaced or contradicted by the strong law of large numbers, must be rejected.

4.5. *The Borel-Cantelli theorems.* These theorems which are mathematically more elementary than the strong law of large numbers again need careful consideration. We begin with the "regular case."

(a) We consider an infinite sequence of events $A_1, A_2, \ldots, A_n, \ldots$ with probabilities $q_1, q_2, \ldots, q_n, \ldots$ for "event" and assume independence and divergence of $\sum_{\nu=1}^{\infty} q_\nu$. Denote by $M_{r,n}$ the set of all sequences in which the event happens at most r times, $0 \leqslant r < n$ out of the first n "trials." This is a set of F (see Chapter II). Since $M_{r,n} \supset M_{r,n+1} \supset \cdots$ this sequence converges as $n \to \infty$. Denote its limit by M_r; this is the set of sequences for which the events happen at most r times. M_r is a Cantor set, as considered in Chapter II; it has a probability in F_1, the field of the extended collective. We wish to show that this probability is zero or, equivalently, that $P(M_r') = 1$, where M_r' is the set of sequences in which more than r events happen.

The probability that no event occurs between the rth and nth trial is

$$(1 - q_r)(1 - q_{r+1}) \cdots (1 - q_n) < e^{-(q_r + q_{r+1} + \cdots + q_n)} \tag{33}$$

and this tends to zero as $n \to \infty$ (r remaining fixed), on account of the divergence of $\Sigma\, q_\nu$. Hence the probability for at least one event after the rth trial is one. Now divide the sequence $\{A_\nu\}$ into two subsequences $\{A_\nu^{(1)}\}$, $\{A_\nu^{(2)}\}$ such that both corresponding series of probabilities diverge. It follows that with probability one at least one event occurs after the rth trial in each of the two subsequences, hence at least two events in all. Continuing in this way we see that the probability is one for the occurrence of more than r events after the rth trial, therefore $P(M_r') = 1$, $P(M_r) = 0$.

Next, consider the sequence of sets: $M_1 \subset M_2 \subset M_3 \subset \cdots$, which converges toward the set M of all sequences which exhibit a finite number of events only. *This set is not in F_1, but in $F_2 - F_1$. It has no probability, but it has measure and we have*

$$\lim_{r\to\infty} P(M_r) = |\lim_{r\to\infty} M_r| = |M|. \tag{34}$$

We have the result.

(a) *Consider a sequence of independent events* A_1, A_2, ..., *with probabilities* q_1, q_2, ..., *where* $\Sigma_{i=1}^{\infty} q_i$ *diverges* (ordinary case). *Then for any* r, *the probability is zero that the event appears at most* r *times*: $P(M_r) = 0$. *In addition the measure of* M *is zero*, i.e., $|M| = 0$, *where* M *is the set of those sequences which exhibit finitely many events only.*[8]

(b) Now we no longer require that the n events be necessarily independent, and we assume that $\Sigma_{i=1}^{\infty} q_i$ converges. Denote by $P_n(x)$ the probability that x out of n events take place. Then from (58) and (59), Chapter IV, we have for the expectation $E[x]$ relative to $P_n(x)$:

$$E[x] = P_n(1) + 2P_n(2) + \cdots + rP_n(r) + \cdots + nP_n(n) = q_1 + q_2 + \cdots + q_n .$$

Since $\Sigma_{i=1}^{\infty} q_i$ converges we have $\Sigma_{i=1}^{\infty} q_i < S$, a constant. Then

$$r[P_n(r) + P_n(r + 1) + \cdots + P_n(n)] < S. \tag{35}$$

The expression in brackets is the probability of at least r out of n successes; we denote it by $W_n(r)$, and thus have, independent of n:

$$W_n(r) < S/r, \qquad r \leqslant n. \tag{35'}$$

Denote by $N_{r,n}$ the set of all sequences with at least r events out of the first n trials. This is a set of F and $N_{r,n} \subset N_{r,n+1} \subset \cdots$. Let $n \to \infty$ and denote by N_r the limit set, i.e., the set of all sequences with at least r events, and by $N_r' = B - N_r$ the part of the set of sequences with less than r events. Consider such a sequence—for example, the sequence 0, 0, 0, ... of *no* event, or a sequence with *one* event at a given place, etc. The set $B - N_r$ contains countably many sequences ("points" in B) only. Each of these sequences has a measure greater than zero; the first one, for example is $(1 - q_1)(1 - q_2)(1 - q_3) \cdots$ in the case of independence, and similarly in the case of dependence (see Chapter IV, Section 8) and this expression is now greater than zero. We have here the exceptional case (irregular valuation) (Chapter II, p. 83), where a single sequence (a single point) has a measure different from zero; such a sequence has no content; all sequences in $B - N_r$ are of this type. Consider also the boundary of N_r. The set N_r is open (it is a sum of basic

[8] Feller [7b], p. 188, says explicitly that for every infinite sequence of results we can establish whether or not infinitely many events occur, and he computes a corresponding "probability." This is contrary to our conception: we think we cannot "establish" whether a sequence contains infinitely many successes; and consequently the set of all sequences with infinitely many successes has measure but not probability. He also formulates (28_2) as a probability statement.

sets) while $B - N_r$ contains no basic set, and we have for the boundary $\mathscr{B}(N_r) = [N_r - N_r] + [(B - N_r) - 0] = B - N_r$, and $| B - N_r | > 0$. Therefore N_r has measure but not probability. [In the non-exceptional case, where Σq_i diverges, each of the sequences of $B - N_r$ has content zero and $| \mathscr{B}(N_r) | = | B - N_r | = 0$.]

From (35′) we have

$$| N_r | < S/r. \tag{35''}$$

Now consider the sequence $N_1 \supset N_2 \supset N_3 \supset \cdots$ which converges toward the set N which contains those and only those sequences with infinitely many events. This set has no content and no verifiable probability and from (35″) its measure is zero. We have

$$| N | = | \lim_{r\to\infty} N_r | = \lim_{r\to\infty} | N_r | = 0. \tag{36}$$

Our result is:

(b) *Consider a sequence of not necessarily independent events with probabilities* $q_1, q_2, \ldots$ *If* Σq_i *converges,* or in other words if the expected value of the number of successes remains finite as $n \to \infty$, *then to any* ϵ *an* r *can be found such that the probability that at least* r *out of* n *events take place is less than* ϵ, *independent of* n. *In addition* $| N | = 0$, *where* N *is the set of those sequences in which infinitely many events appear.*

We may unite parts of statements (a) and (b) by saying:

Let the events be independent and denote by M *the set of sequences with a finite number of events. Then* $| M | = 1$ *if* Σq_i *converges and* $| M | = 0$ *if this sum diverges.* Of course, if N is the set of sequences with infinitely many events each, then $| N | = 0$ in the first and $| N | = 1$ in the second case.

Problem 12. Two gamblers A and B toss a coin with probability p for heads and q for tails. Let x be the number of heads and $y = n - x$ that of tails in n trials. A wins if $x \geqslant 6$ and $x = y + 2$, and similarly for B (this is a crude model of the rules of playing a "set" of tennis). Prove that with probability one the game will finish for a finite n and that this is a probability and not only a measure. Can you make the same statement if the difference between x and y is to equal $d > 2$?

B. Characteristic Function (Sections 5-8)

5. The Characteristic Function

5.1. *Various definitions.* What is known today as the characteristic function of a distribution or, with a slightly different definition, as the generating function of a distribution was essentially introduced by Laplace into probability calculus, and it has proved to be a mathematical tool of great usefulness. The theory of the Laplace transform is a well-developed domain of analysis successfully employed in several branches of applied mathematics.

Let $p(x)$, $x = 0, 1, 2, \ldots$ be an arithmetic distribution. If the series $F(t) = \sum_{x=0}^{\infty} p(x)t^x$ converges, for some t, where $|t| < t_0$, then $F(t)$ is called the *generating function* of $p(x)$. Here t is, in general, a complex variable. The term "generating function" is also used when the coefficients of the powers of t are real numbers, not necessarily probabilities, and $\sum_{\nu=0}^{\infty} a_\nu t^\nu$ converges in some interval $|t| < t_0$. The series is then the generating function of the sequence $\{a_\nu\}$.

In this book the generating function has already been used in our study of the binomial distribution in Chapter IV. There, Eq. (13′), the expression $F(t) = p_n(0) + p_n(1)t + \cdots + p_n(n)t^n$ served to compute the moments of the binomial distribution, and using the definition $p_n(x) = \binom{n}{x}q^x p^{n-x}$, $x = 0, 1, \ldots, n$ we saw that $\sum_{x=0}^{n} p_n(x)t^x = (p + qt)^n$. Since for $n = 1$ the binomial distribution reduces to the simple alternative $p_1(0) = p$, $p_1(1) = q$, whose generating function is $p + qt$, the above equation is equivalent to the statement: the generating function of the distribution of the sum of n chance variables, each of which is subject to the same simple alternative, equals the nth power of the generating function of the alternative. We shall see, presently (Section 5.3), that this is a special case of an important theorem on generating functions.

When the coefficients are probabilities, the generating function $F(t)$ of a distribution with the cumulative distribution function $P(x)$ is defined by

$$F(t) = \int t^x \, dP(x). \tag{37}$$

This formula is equivalent to $F(t) = \sum_x t^x p(x)$ for a discrete, and to $F(t) = \int t^x p(x)\, dx$ for a continuous distribution. We could also define $F(t)$ by

$$F(t) = E[t^x], \qquad \text{relative to} \quad P(x). \tag{37′}$$

In general, we have to subject $P(x)$ to suitable conditions in order to make sure that the integral in (37) exists. This can be avoided by restricting the range of the variable t. Let us introduce a real variable u by

$$t = e^{iu} = \cos u + i \sin u. \tag{38}$$

In this way, t is restricted to the points on the circumference of the unit circle in the complex t-plane and $|t| = 1$. $F(t)$ now becomes

$$F(t) = \int e^{ixu}\, dP(x) = f(u), \tag{39}$$

where $f(u)$ is a complex-valued function of the real variable u. Again, $f(u)$ equals the expectation of e^{ixu} relative to $P(x)$. Since $|e^{ixu}| = 1$, one has

$$|f(u)| = \left| \int e^{ixu}\, dP(x) \right| \leqslant \int |e^{ixu}|\, dP(x) = \int dP(x) = 1, \tag{40}$$

that is, the integral that defines $f(u)$ exists and converges absolutely and uniformly in u. The function $f(u)$, which will be used extensively in the following, is called the *characteristic function* (c.f.) of the distribution. The reader may keep in mind that u is a real variable and $f(u)$ nothing but an abbreviation for the pair of functions

$$\int \cos xu\, dP(x), \qquad \int \sin xu\, dP(x). \tag{41}$$

This abbreviation leads to various computational simplifications. We show that $f(u)$ *is uniformly continuous for all* u. Consider $f(u+h) - f(u) = \int e^{iux}(e^{ihx} - 1)\, dP(x)$; then $|f(u+h) - f(u)| \leqslant \int |e^{ihx} - 1|\, dP(x)$. Now, let ϵ be an arbitrary number. We take X so large that $\int_{|x|>X} dP(x) < \epsilon/4$ and h so small that for $|x| < X$, $|\epsilon^{ixh} - 1| < \epsilon/2$. Then, $|f(u+h) - f(u)| \leqslant \int_{-X}^{X} |\epsilon^{ixh} - 1|\, dP(x) + 2\int_{|x|\geqslant X} dP(x) \leqslant \epsilon$. Also from (40): $f(u)| \leqslant 1$, for all u.

It is also easily seen that always $f(-u) = \overline{f(u)}$.

If $P(x)$ is a continuous distribution with density $p(x) = P'(x)$, Eq. (37) or (39) can be interpreted as an integral transformation of the function $p(x) = dP(x)/dx$. The function

$$\frac{1}{\pi} f(u) = \frac{1}{\pi} \int e^{ixu}\, p(x)\, dx \tag{42}$$

is commonly called the *Fourier transform* of the function $p(x)$. Equation (37) with $t = e^{-s}$ is the so-called *Laplace transform* of $p(x)$:

$$F(t) = L(s) = \int e^{-sx} p(x)\, dx\,. \tag{43}$$

If, in (37), t is replaced by e^{u}, one obtains the moment generating function

$$f^*(u) = \int e^{xu}\, dP(x)\,. \tag{44}$$

If the integral in (44) exists, we may differentiate under the integral sign provided that the resulting expressions exist and are uniformly convergent. This gives

$$\frac{d^\nu f^*(u)}{du^\nu} = \int x^\nu e^{xu}\, dP(x)\,. \tag{44'}$$

If then u is set equal to zero, one obtains the zero moment of νth order,

$$M_\nu^{(0)} = \frac{d^\nu f^*(u)}{du^\nu}\bigg]_{u=0}\,. \tag{44''}$$

Clearly, all the functions $f(u)$, $f^*(u)$, $L(s)$ can be regarded as special cases of the generating function $F(t)$.

Let us return to the characteristic function $f(u)$. We assume that for $P(x)$ the absolute moments of any order exist, and find by differentiation of (39) with $d^\nu f(u)/du^\nu = f^{(\nu)}(u)$:

$$f^{(\nu)}(u) = i^\nu \int x^\nu e^{iux}\, dP(x). \tag{45}$$

Here, $| \int x^\nu e^{iux}\, dP(x) | \leqslant \int | x^\nu |\, dP(x)$, and we thus see that the integral in (45) exists; by putting $u = 0$, we find

$$f^{(\nu)}(0) = i^\nu \int x^\nu\, dP(x) = i^\nu M_\nu^{(0)},$$

or

$$M_\nu^{(0)} = i^{-\nu} \frac{d^\nu f(u)}{du^\nu}\bigg]_{u=0}\,. \tag{45'}$$

We may expand $f(u)$ in a MacLaurin series in the neighborhood of $u = 0$ and obtain[1]

$$f(u) = 1 + \sum_{\nu=1}^{n} \frac{M_\nu}{\nu!}(iu)^\nu + o(u^n), \tag{46}$$

where the error term divided by u^n tends to zero as $u \to 0$.[2] From (46) and (45) we see that $f(u)$ is also a moment generating function.

5.2. *Examples.* We consider now some examples of characteristic functions.

(1) *For the normal distribution* $p(x) = (\sqrt{2\pi}s)^{-1} \exp[-(x - a)^2/2s^2]$, the c.f. is $f(u) = (\sqrt{2\pi}s)^{-1} \int \exp[iux - (x - a)^2/2s^2]\, dx$. By the substitution

$$z = \frac{x - a}{s} - ius$$

[1] A proof of this expansion is in Chapter VI, p. 265.

[2] We use here some well-known notation: $f(x) = O[g(x)]$ means that $f(x)/g(x)$ is bounded as x tends toward its limit; $f(x) = o[g(x)]$ means that the quotient tends toward zero as x tends toward its limit; $f(x) \sim g(x)$ means that in the limit the above quotient tends toward one.

we obtain

$$f(u) = e^{iau-(s^2u^2/2)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty-ius}^{\infty-ius} e^{-z^2/2}\, dz.$$

It is known that for any real a,

$$\int_{-\infty-ia}^{\infty-ia} e^{-z^2/2}\, dz = \sqrt{2\pi}\,.$$

Hence

$$f(u) = e^{iau-(s^2u^2/2)}\,. \tag{47}$$

If $a = 0$, $s = 1$, i.e., for $\phi(x) = (1/\sqrt{2\pi})\, e^{-x^2/2}$, the c.f. is simply $f(u) = e^{-u^2/2}$ (see Problem 14.) The moments of the normal distributions can now be easily computed. More follows in Section 7.2.

(2) *For the Poisson distribution*

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

$$f(u) = e^{-a} \sum_{x=0}^{\infty} e^{iux} \frac{a^x}{x!} = e^{-a} \sum_{x=0}^{\infty} \frac{(ae^{iu})^x}{x!} = e^{-a+ae^{iu}} = e^{a(e^{iu}-1)}\,. \tag{48}$$

We note that as $u \to \infty$ the c.f. of the density in example (1) tends to zero, while that of the arithmetic distribution in example (2) is periodic and does not tend to zero (or to any limit) as $u \to \infty$.

(3) From the results of Chapter IV, p. 170, or by direct computation we find the *c.f. of the binomial distribution*

$$f(u) = (p + qe^{iu})^n. \tag{49}$$

(4) Consider the *Cauchy distribution*

$$p(x) = \frac{1}{\pi} \frac{a}{a^2 + (x-b)^2}\,, \qquad -\infty < x < +\infty, \quad a > 0. \tag{50}$$

It is seen that $\int p(x)\, dx = 1$, but that neither the mean value nor the moments of higher order exist, since the integrals $\int x\, dx/(a^2 + x^2)$, $\int x^2\, dx/(a^2 + x^2)$, ... diverge. But the c.f. $f(u)$ exists and equals

$$f(u) = e^{biu-a|u|}\,. \tag{50'}$$

The form of (50′) shows that the product of two characteristic functions $f_i(u)$, $i = 1, 2$, is a c.f. of the same type with parameters $a_1 + a_2$ and

$b_1 + b_2$. Hence, anticipating the uniqueness theorem of Section 6.2, we recover the result of Problem 7 (p. 233) [in a slightly generalized form since (50) contains two parameters].

(5) If the density $p(x)$ is the "gamma distribution"

$$p(x) = \frac{1}{m!} x^m e^{-x}, \qquad x > 0, \qquad \text{otherwise zero,} \tag{51}$$

where $m! = \Gamma(m+1)$; the reader will verify (Problem 17) that the c.f. equals

$$f(u) = (1 - iu)^{-(m+1)}. \tag{51'}$$

The product of two characteristic functions with parameters $m_1 + 1 = \lambda_1$, $m_2 + 1 = \lambda_2$ is of the same form with parameter $\lambda_1 + \lambda_2$. This generalizes to the product of n characteristic functions of the form (52).

If we set

$$p(x) = \frac{1}{2^{m/2}\Gamma\left(\frac{m}{2}\right)} x^{\frac{m}{2} - 1 - \frac{x}{2}} \qquad \text{(Helmert's distribution)} \tag{52}$$

the characteristic function is

$$f(u) = (1 - 2iu)^{-m/2} \tag{52'}$$

(see computation Chapter VIII, Eq. (68)).

5.3. *Generating function of convolution.* Returning to our problem of independent chance variables, we form the generating functions for each of the given individual distributions $P_\nu(x)$:

$$F_\nu(t) = \int^x t \, dP_\nu(x), \qquad \nu = 1, 2, \ldots, n. \tag{53}$$

And we denote by $G_n(t)$ the generating function of $Q_n(x) = \Pr\{x_1 + x_2 + \cdots + x_n \leqslant x\}$:

$$G_n(t) = \int t^x \, dQ_n(x). \tag{54}$$

Introducing $dQ_n(x)$ from (24'), we obtain, with $x = y + z$

$$G_n(t) = \int t^x \int dQ_{n-1}(x - z) \, dP_n(z) = \iint t^{y+z} \, dQ_{n-1}(y) \, dP_n(z)$$

$$= \int t^y \, dQ_{n-1}(y) \cdot \int t^z \, dP_n(z) = G_{n-1}(t) \cdot F_n(t).$$

Since $G_1(t) = F_1(t)$ because $Q_1(x) = P_1(x)$, Eq. (54) implies

$$G_n(t) = F_1(t) \cdot F_2(t) \cdots F_n(t). \tag{55}$$

We shall refer to this important relation as the *multiplication theorem*: *The generating function of the distribution of a sum of independent chance variables equals the product of the generating functions of the individual distributions.* In the case of discrete distributions, this result reduces to the [$(n - 1)$ times repeated] multiplication of power series

$$\sum a_\nu t^\nu \cdot \sum b_\mu t^\mu = \sum c_\rho t^\rho, \qquad \text{where} \qquad c_\rho = a_0 b_\rho + a_1 b_{\rho-1} + \cdots + a_\rho b_0$$

is the *convolution* of the sequences $\{a_\nu\}$ and $\{b_\nu\}$ (p. 226).

The relation (55) applies to all special cases of generating functions, for example, to the characteristic functions. If $f_\nu(u) = \int e^{iux}\, dP_\nu(x)$, $\nu = 1, 2, \ldots, n$, $g_n(u) = \int e^{iux}\, dQ_n(x)$, Q_n convolution of the f_ν, then:

$$g_n(u) = f_1(u) \cdot f_2(u) \cdot \quad \cdots \quad \cdot f_n(u). \tag{55'}$$

The multiplication theorem permits the computation of $g_n(u)$ from the given $P_\nu(x)$. However, the problem of determining $Q_n(x)$ from $g_n(u)$, the *inversion* of the relation between characteristic function and distribution requires additional analysis.

Before turning to the problems of inversion, we consider the following simple *applications of the multiplication rule.*

(1) We have seen that the binomial distribution is obtained by n-fold convolution from the simple alternative $p(0) = p$, $p(1) = q$, $p + q = 1$. The generating function of this alternative is $p + qt$, hence, that of the binomial distribution, $q_n(x) = \binom{n}{x} q^x p^{n-x}$, is $F(t) = (p + qt)^n$, as found before. By the rule for finding the moments we obtain $F'(1) = nq(q+p)^{n-1} = nq$, the mean value; also $F''(1) = n(n-1)q^2(q+p)^{n-2} = n(n - 1)\, q^2 = \Sigma_{x=0}^{n} x(x - 1)\, p_n(x)$. Hence, the variance is $s_n{}^2 = n(n - 1)\, q^2 + nq - n^2q^2 = npq$, etc., as in Chapter IV.

(2) We have seen on p. 227 that the negative binomial distribution is obtained by n-fold convolution from the geometric distribution qp^x, $x = 0, 1, \ldots$. Its generating function $F(t)$ is $\Sigma_{x=0}^{\infty} q(tp)^x = q/(1 - tp)$. The generating function of the negative binomial distribution is by the multiplication theorem

$$g_n(t) = \left(\frac{q}{1 - tp}\right)^n. \tag{56}$$

Let us find mean value and variance. From $F(t) = q(1 - tp)^{-1}$, $F'(t) = pq(1 - tp)^{-2}$, $F''(t) = 2p^2q(1 - tp)^{-3}$, we find $F(1) = 1$,

$F'(1) = p/q$, $F''(1) = 2p^2/q^2$, $F''(1)+F'(1)-F'(1)^2 = (2p^2/q^2) + (p/q) - (p^2/q^2) = p/q^2$. Using the rules (13) and (17) we have for the negative binomial distribution

$$b_n = \frac{np}{q}, \qquad s_n{}^2 = \frac{np}{q^2}. \tag{57}$$

Formulas (56) and (57) also hold true for n positive and not necessarily integral.

Problem 13. Compute the characteristic functions $f(u)$

(a) for the uniform distribution between $x = -c$ and $x = c$;

(b) for the distribution with the density $p(x) = ce^{-cx}$ between 0 and ∞. Plot graphs for (a) and (b).

Problem 14. Prove that if $f(u)$ is the c.f. for a chance variable t, the c.f. for $t + a$ is $f(u) \cdot e^{iau}$, and that for bt is $f(bu)$. Give the corresponding formula for the chance variable $a + bt$, and check by means of the result (47). [Note that if $Pr\{t \leqslant x\} = P(x)$, then $\Pr\{bt + a \leqslant x\} = P((x - a)/b)$.]

Problem 15. If a distribution has mean value a and variance s^2 (or: if the random variable ξ has mean value a and standard deviation s) we call $\eta = (\xi - a)/s$ the *standardized variable*; it has mean value 0 and standard deviation 1. Show that in the case of the binomial distribution, the characteristic function $f(u)$ of the standardized variable $\eta = (\xi - nq)/\sqrt{npq}$ is given by

$$f(u) = (pe^{-iu\sqrt{q/np}} + qe^{iu\sqrt{p/nq}})^n.$$

Problem 16. Derive the formulas

$$b_n = a_1 + a_2 + \cdots + a_n$$
$$s_n{}^2 = r_1{}^2 + r_2{}^2 + \cdots + r_n{}^2$$

from the multiplication theorem (55').

Problem 17. Compute the c.f. for the distribution with the density $p(x) = x^m e^{-x}/m!$, $x \geqslant 0$.

Problem 18. Prove: if a distribution is symmetric and continuous, $P(x) = 1 - P(-x)$, its c.f. is real.

6. Inversion

6.1. *Arithmetic distribution and density.* We start the discussion of the inversion problem by considering $P(x)$ to be a step function, the steps of which coincide with the integers 0, 1, 2, ... If $p(\nu)$ denotes the magnitude of the saltus of $P(x)$ at $x = \nu$, (i.e., the probability of the label value ν), the c.f. of P can be written as

$$f(u) = \int e^{ixu}\, dP(x) = \sum_{x=0}^{\infty} e^{ixu}\, p(x) = p(0) + p(1)\cos u + p(2)\cos 2u + \cdots + i[p(1)\sin u + p(2)\sin 2u + \cdots]; \qquad (58)$$

The determination of $p(x)$ from $f(u)$ can be carried out in the same way as the determination of the coefficients of the Fourier expression of a given function. Multiply both sides of Eq. (58) by $\cos yu$, where y is one of the numbers 0, 1, 2, ... and integrate term by term from $-\pi$ to π:

$$\int_{-\pi}^{\pi} f(u)\cos yu\, du = p(0)\int_{-\pi}^{\pi} \cos yu\, du + p(1)\int_{-\pi}^{\pi} \cos u \cos yu\, du + \cdots + i\left[p(1)\int_{-\pi}^{\pi} \sin u \cos yu\, du + \cdots\right]. \qquad (59)$$

The values of the integrals on the right-hand side of this equation are

$$\begin{aligned} &\int_{-\pi}^{\pi} \sin nx \cos mx\, dx = 0, &&\text{for all integers } n \text{ and } m,\\ &\int_{-\pi}^{\pi} \cos nx \cos mx\, dx \begin{cases} = 0, & \text{if } n \neq m\\ = \pi, & \text{if } n = m.\end{cases} \end{aligned} \qquad (60)$$

Therefore, (58) leads to

$$p(y) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(u)\cos yu\, du. \qquad (61)$$

Multiplying Eq. (58) by $\sin yu$, one obtains

$$p(y) = \frac{1}{i\pi}\int_{-\pi}^{\pi} f(u)\sin yu\, du = -\frac{1}{\pi}\int_{-\pi}^{\pi} f(u) i \sin yu\, du. \qquad (61')$$

Adding (61) and (61′), writing x instead of y and e^{-ixu} for $\cos xu - i \sin xu$, we arrive at the inversion formula of (39):

$$p(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(u) e^{-ixu}\, du. \qquad (62)$$

Formula (62) has been proved here under the assumption of an arithmetic distribution $P(x)$ with saltus $p(x)$ at $x = 0, 1, 2, \ldots$. A similar formula holds for a continuous distribution $P(x)$ with a density $p(x) = dP(x)/dx$. We remember Eq. (42) and assume that $p(x)$ is of bounded variation in every finite interval. Then we obtain from Fourier's theorem which we do not derive here, that

$$p(x) = \lim_{\lambda\to\infty} \frac{1}{2\pi} \int_{-\lambda}^{\lambda} f(u)e^{-iux}\, du \tag{62'}$$

at every point of continuity of $p(x)$. If $p(x)$ has a jump at x the left-hand side is replaced by $\frac{1}{2}[p(x+0) + p(x-0)]$. If $|f(u)|$ is absolutely integrable in $-\infty$, $+\infty$, we may replace $\lim_{\lambda\to\infty} \int_{-\lambda}^{\lambda}$ by $\int_{-\infty}^{\infty}$.

We have seen in (45′) that the differentiability properties of $f(u)$ are related to the existence of the moments of $P(x)$, hence to the behavior of $P(x)$ for large values of X. Likewise, the behavior of $f(u)$ at infinity is related to the differentiability of $P(x)$. If $P'(x) = p(x)$ exists everywhere and is continuous, the integral (42) exists and $f(u) \to 0$ as $|u| \to \infty$. If in addition $p^{(n)}(x)$ exists for all x and $|p^{(n)}(x)|$ is integrable in $-\infty$, $+\infty$, repeated integration by parts shows that

$$|f(u)| < \frac{K}{|u|^n} \qquad \text{for all} \quad u.$$

Hence, if $|p^{(2)}(x)|$ is integrable in $-\infty$, $+\infty$, then $|f(u)|$ is integrable there and (62′) may be replaced by

$$p(x) = \frac{1}{2\pi} \int f(u)e^{-iux}\, du. \tag{62''}$$

6.2. General inversion. Uniqueness theorem. We consider a general distribution $P(x)$.[1]

Let $f(u)$ be defined as before by

$$f(u) = \int e^{iux}\, dP(x). \tag{63}$$

Our contention is that the inversion of (63) for any x at which $P(x)$ has no step discontinuities reads

$$P(x) = \frac{1}{2\pi} \lim_{U\to\infty} \int_{-U}^{U} \frac{1 - e^{-ixu}}{iu} f(u)\, du + \text{constant}. \tag{64}$$

The constant on the right-hand side will be determined by the condition that $\lim_{x\to-\infty} P(x) = 0$. More precisely, we want to prove:

[1] See Lévy [19a], p. 166 ff; and Cramér [5], pp. 28–29. See v. Mises, "Fundamentalsätze." *Math. Z.* **4** (1919), p. 34.

If $f(u)$ is the characteristic function of a distribution $P(x)$ and x as well as the origin are continuity points of $P(x)$ then (64) *holds with the constant* $= -P(0)$.

To prove (64), we replace $f(u)$ in (64) by its definition (63) calling the (real) integration variable z instead of x. In this way, the integral in (64) which we abbreviate by $R \equiv \lim_{U\to\infty} R_U$ becomes a double integral

$$R = \frac{1}{2\pi} \lim_{U\to\infty} \int_{-U}^{U} du \int \frac{1 - e^{-ixu}}{iu} e^{izu}\, dP(z). \tag{65}$$

Here, one may reverse the order of integration since the integral converges absolutely with respect to z, and with respect to u the limits are finite. Hence,

$$R = \frac{1}{2\pi} \lim_{U\to\infty} \int \left[\int_{-U}^{U} \frac{1 - e^{-iux}}{iu} e^{iuz}\, du \right] dP(z). \tag{65'}$$

Now,

$$\frac{e^{izu} - e^{i(z-x)u}}{iu} = \frac{\cos zu - \cos (z - x)\, u}{iu} + \frac{\sin zu - \sin (z - x)\, u}{u}, \tag{66}$$

being thus represented as the sum of an odd and an even function of the variable u. In the integration with respect to u from $-U$ to U the contributions of the odd term cancel, the contributions of the even term are equal for the segments $(-U, 0)$ and $(0, U)$. The expression R_U may thus be written

$$R_U = \frac{1}{\pi} \int \left[\int_0^U \frac{\sin zu - \sin(z - x)\, u}{u}\, du \right] dP(z). \tag{67}$$

In (67) the two integrals in brackets are, for constant x, functions of z. We write

$$I_U(z) = \int_0^U \frac{\sin zu}{u}\, du, \qquad I(z) = \lim_{U\to\infty} I_U(z). \tag{68}$$

The function $I(z)$ is known as *Dirichlet's integral* or Dirichlet's discontinuity factor. We see that $\int_0^U (\sin zu)/u\, du = \int_0^{zU} (\sin u)/u\, du$. It is proved in calculus that $\int_0^x (\sin t)/t\, dt$ is bounded for all $x > 0$ and tends to $\pi/2$ as $x \to \infty$. Therefore,

$$I(z) = \frac{\pi}{2} \quad \text{for} \quad z > 0, \qquad = 0 \quad \text{for} \quad z = 0, \qquad = -\frac{\pi}{2} \quad \text{for} \quad z < 0. \tag{69}$$

The convergence of $I_U(z)$ to $I(z)$ is uniform with respect to z for $z > \delta > 0$ and for $z < -\delta$, while for $|z| \leqslant \delta$ and all U:

$$\left| \frac{1}{\pi} \int_0^U \frac{\sin zu}{u}\, du \right| < 1. \tag{70}$$

Thus, from (69) $I(z)$ is a unit-step function that jumps from $-\pi/2$ to $\pi/2$ at $z = 0$. The function $I(z - x)$, considered as a function of z for constant x, has its step at $z = x$.

According to (67) and (68) we have to consider the integral

$$R_U = \frac{1}{\pi} \int [I_U(z) - I_U(z - x)]\, dP(z). \tag{71}$$

Assume $x > 0$ and divide the interval of integration into the intervals $(-\infty, -\delta)$, $(-\delta, +\delta)$, $(+\delta, x - \delta)$, $(x-\delta, x + \delta)$, and $(x + \delta, +\infty)$, where δ is a positive number less than $x/2$. In the first interval $z < -\delta$ (and thus $z - x < -\delta$), the difference in the bracket in (71) and therefore the integral tends uniformly to zero as $U \to \infty$; the same consideration applies to the last interval where $z > x + \delta$. Next, the integral from $-\delta$ to $+\delta$, and similarly that from $x - \delta$ to $x + \delta$, tend to zero as $\delta \to 0$, by the estimate (70) and the assumed continuity of P at the points 0 and x. In the middle interval from δ to $x - \delta$ the difference in bracket tends uniformly to π as $U \to \infty$, and therefore the integral tends toward $P(x - \delta) - P(\delta)$ and toward $P(x) - P(0)$ as $\delta \to 0$. Thus,

$$\lim_{U\to\infty} R_U = P(x) - P(0), \qquad x > 0.$$

Completely analogous reasoning applies if $x < 0$. Then, in the interval $(x + \delta, -\delta)$ the value in bracket tends to $-\pi$ as $U \to \infty$ and the integral toward $- \mid P(-\delta) - P(x + \delta) \mid$, which tends again toward $P(x) - P(0)$ as $\delta \to 0$. Thus, (64) is proved and the constant found to be $-P(0)$. If we wish an arbitrary continuity point α of $P(x)$ to take the place of the origin, Eq. (64) is replaced by

$$P(x) - P(\alpha) = \frac{1}{2\pi} \lim_{U\to\infty} \int_{-U}^{U} \frac{e^{-iu\alpha} - e^{-iux}}{iu} f(u)\, du. \tag{64'}$$

When $P(x)$ is the integral of a density $p(x)$, we recover the inversion formula (62′) from (64) by formal differentiation.

From (64′) we obtain for every continuity point of $P(x)$, letting $\alpha \to -\infty$:

$$P(x) = \frac{1}{2\pi} \lim_{\alpha\to-\infty} \lim_{U\to+\infty} \int_{-U}^{U} \frac{e^{-iu\alpha} - e^{-iux}}{iu} f(u)\, du. \tag{64''}$$

Here the limit with respect to α is taken through continuity points of $P(x)$. Hence we obtain the *uniqueness theorem*:

The distribution function $P(x)$ is uniquely determined by its characteristic function $f(u)$.

7. Solution of the Summation Problem. Stability of the Normal Distribution and of the Poisson Distribution

7.1. *Application to the summation problem.* In the second section of this chapter the distribution $Q_n(x)$ of the sum of n independent chance variables was found by means of a recursion formula. The results obtained in the last two sections make it possible now to write down $Q_n(x)$ for any n in explicit form.

Let $P_\nu(x)$ be the given distributions of the chance variables x_ν. The characteristic function of the distribution $P_\nu(x)$ is then $f_\nu(u) = \int e^{ixu}\, dP_\nu(x)$. According to the multiplication theorem, the characteristic function of the distribution $Q_n(x)$ of the sum of the chance variables $x_1 + x_2 + \cdots + x_n$ was seen to be the product $f_1(u) f_2(u) \cdots f_n(u)$. The inversion theorem (64) then gives the distribution $Q_n(x)$ in the form

$$\begin{aligned} Q_n(x) - Q_n(0) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1 - e^{ixu}}{iu} f_1(u) f_2(u) \cdots f_n(u)\, du \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1 - e^{ixu}}{iu} \left\{ \prod_{\nu=1}^{n} \int e^{ixu}\, dP_\nu(x) \right\} du. \end{aligned} \tag{72}$$

Although this formula offers the complete and explicit solution of the summation problem, it will not, in general, be very helpful if n is large, since the actual computation of the product of n functions may be too cumbersome. In the next chapter, we will develop an approximation for $Q_n(x)$ which, if n is large, gives what is called an "asymptotic expression" for $Q_n(x)$, without requiring the multiplication of n factors. Meanwhile we shall derive from our results some important conclusions concerning the cases where all $P_\nu(x)$ are *normal Gaussian distributions*, or all *Poisson distributions*.

7.2. *Stability of the normal distribution.* We have previously found, in Eq. (47), that the c.f. of a normal distribution with mean a and variance r^2,

$$p(x) = \frac{1}{\sqrt{2\pi} r} \exp\left(- \frac{(x - a)^2}{2r^2}\right)$$

equals

$$f(u) = \exp\left(- \frac{r^2}{2} u^2 + iau\right). \tag{73}$$

Suppose now that a set of normal distributions $N(a_1, r_1{}^2)$, $N(a_2, r_2{}^2)$, ..., $N(a_n, r_n{}^2)$ is given. The c.f. of $N(a_\nu, r_\nu{}^2)$ is according to (73)

$$f_\nu(u) = e^{ia_\nu u - \frac{1}{2}(u^2 r_\nu{}^2)}.$$

The characteristic function belonging to the distribution of the sum of the n independent chance variables is, according to the multiplication theorem, the product

$$g_n(u) = f_1(u)f_2(u) \cdots f_n(u) = e^{ib_n u - \frac{1}{2}(u^2 s_n^2)}, \tag{74}$$

where s_n^2 has been written for $r_1^2 + r_2^2 + \cdots + r_n^2$ and b_n for $a_1 + a_2 + \cdots + a_n$. This product $g_n(u)$ is again of the same form as (73) and, therefore, it is the c.f. of the normal distribution $N(b_n, s_n^2)$, since we know from the uniqueness theorem that a distribution is uniquely determined by its characteristic function. The result is *that the sum of n independent normally distributed chance variables is also normally distributed.* At the same time, the summation theorems of mean value and variance for a sum of chance variables have been confirmed for the special case under consideration.

The above result states that normal distributions reproduce themselves in the process of summation of chance variables. This property is sometimes termed the *stability* of the normal distribution.

7.3. *Reproductive property of the Poisson distribution. Converse theorems.* Next consider n independent chance variables, each subject to a Poisson distribution

$$p_\nu(x) = \frac{a_\nu^x}{x!} e^{-a_\nu}, \qquad \nu = 1, \ldots, n. \tag{75}$$

From (48) we know the corresponding characteristic functions:

$$f_\nu(u) = e^{a_\nu(e^{iu}-1)}, \qquad \nu = 1, \ldots, n. \tag{76}$$

We consider the characteristic function belonging to the sum of n independent chance variables, viz.,

$$\begin{aligned} g_n(u) &= f_1(u)f_2(u) \cdots f_n(u) \\ &= \exp\,[(e^{iu} - 1)(a_1 + a_2 + \cdots + a_n)] = \exp\,[(e^{iu} - 1)\, b_n], \end{aligned} \tag{77}$$

which is seen to be again the characteristic function of a Poisson law. Therefore, by the uniqueness theorem, we conclude that *the sum of n independent chance variables, each subject to Poisson's law, is again subject to Poisson's law.* Of course the mean value is b_n.

We mention without proofs that certain converses of the last two theorems are valid. Cramér[1] has proved: *If the sum of n independent*

[1] Cramér [5], p. 52; the original paper is "Ueber eine Eigenschaft der normalen Verteilungsfunktion." *Math. Z.* **41** (1936), pp. 405–414.

random variables is normally distributed, then each of the n variables is itself normally distributed. Thus, a normal distribution can never be the *exact* distribution of the sum of n independent non-normal component-variables.

D. A. Raikow proved the analogous fact for the Poisson distribution: *If the sum of n independent chance variables has the Poisson distribution, then for each term the Poisson distribution holds.*

We have found in (50′) that the characteristic function of the Cauchy distribution with density

$$p_{a,b}(x) = \frac{1}{\pi} \frac{a}{a^2 + (x - b)^2} \qquad \text{is} \qquad f_{a,b}(u) = e^{biu - a|u|}.$$

Thus $f_{a_1,b_1} \cdot f_{a_2,b_2} = f_{a_1+a_2,b_1+b_2}$; it is thus seen that the Cauchy distribution also reproduces itself. The same is true of Helmert's distribution (52).

8. Continuity Theorem for Characteristic Functions

8.1. *Statement of the theorem.* We have seen in Section 6 that there is a one-to-one relation between a distribution $P(x)$ and its characteristic function $f(u)$. We shall need an even more more far-reaching statement, namely, that under certain quite general conditions, the relations $P_n(x) \to P(x)$ and $f_n(u) \to f(u)$ are equivalent.[1] The proof of this continuity theorem, as it is called, uses as an essential tool two theorems of analysis due to E. Helly, which will be designated by (A) and (B); they will be formulated and used in our proof of the continuity theorems but not explicitly proved in this text.[2] The *continuity theorem* consists of two parts:

(1) *Given a sequence of distributions $P_1(x)$, $P_2(x)$, ..., $P_n(x)$, ... which converge toward a distribution function $P(x)$ in every continuity point of $P(x)$; then, the sequence of the corresponding characteristic functions $f_1(u)$, $f_2(u)$, ..., $f_n(u)$, ... converges toward the characteristic function $f(u)$ of $P(x)$.*

[1] The theorem has been proved under not very different conditions by P. Lévy [19a], pp. 195–197; by V. Glivenko, "Sul teorema limite della teoria delle funzioni caratteristiche." *Giorn. Ist. Ital. Attuari* 7 (1936), pp. 160–167, generalized by P. Lévy, [19b], p. 49; and by Cramér [5], p. 29. See also the proof in Gnedenko [10], p. 208.

[2] E. Helly, "Über lineare Funktionaloperationen." *Sitzungsberichte Akad. Wiss. Wien Math. Nat. Kl.* 121, Abt. IIa (1912), pp. 1–33. For proofs see Cramér [4], p. 60 ff, p. 74 ff, and Gnedenko [10], p. 204 ff, p. 207 ff; E. Lukacs, *Characteristic Functions*, p. 49 ff. New York, 1960.

More important for us is the second theorem:

(2) *If the sequence of characteristic functions* $f_1(u)$, $f_2(u)$, ..., $f_n(u)$, ... *converges toward a continuous function* $f(u)$ *then the sequence of the corresponding distributions* $P_1(x)$, $P_2(x)$, ..., $P_n(x)$, ... *converges toward a distribution* $P(x)$, *which, on account of* (1), *has* $f(u)$ *as characteristic function.*

8.2. *Statement of two theorems by Helly and proof of continuity theorem* (1). To prove (1), we use the following theorem of Helly.[3]

(A) *Consider a sequence of distribution functions which converge to a non-decreasing function* $P(x)$, *in every continuity point of* $P(x)$[4] *and also at* $x = -\infty$, $x = +\infty$. *Let* $g(x)$ *be everywhere continuous*; *then*

$$\lim_{n\to\infty} \int g(x)\, dP_n(x) = \int g(x)\, dP(x). \tag{78}$$

[Here $P(x)$ is not necessarily a distribution. This possibility will be illustrated by an example in Section 8.4.]

The proof of (1) is immediate, if the validity of (A) is assumed. In fact, with $g(x) = e^{iux}$, which is continuous on the whole real axis, (78) shows directly that $f_n(u) \to f(u)$ and part (1) of the continuity theorem is proved.

To prove (2), we need the second theorem of Helly:

(B) *Any sequence of uniformly bounded, non-decreasing functions* $P_1(x)$, $P_2(x)$, ..., $P_n(x)$, ... *contains a subsequence* $P_{n_1}(x)$, $P_{n_2}(x)$, ..., $P_{n_\nu}(x)$, ... *which converges toward a nondecreasing function* $P(x)$ *in every continuity point of* $P(x)$.

8.3. *Proof of continuity theorem* (2) *by means of Helly's theorem B.* We apply this theorem to the sequence $\{P_n(x)\}$ of distributions considered in (2) of Sect. 8.1 and conclude that it contains a subsequence $\{P_{n_\nu}(x)\}$ possessing the above property; we can assume that $P(x)$ (which is non-decreasing) is continuous to the right. First, one has to prove that $P(x)$ is a distribution, a fact that is by no means obvious (see Section 8.4).

Since obviously $0 \leqslant P(x) \leqslant 1$, we have only to show that $P(-\infty) = 0$, and $P(+\infty) = 1$. If this were not so, the difference $\delta = P(+\infty) - P(-\infty)$ would be <1, viz., $\delta < 1$; next choose a positive $\epsilon < 1 - \delta$; since the characteristic

[3] Theorem (*A*) is often designated as an "Extension of Helly's second theorem," the latter relating to a finite interval $[a, b]$.

[4] A sequence of functions which converges to a limit $f(x)$ at every continuity point of $f(x)$ is often said to *converge weakly* to $f(x)$. It may be shown that from "convergence in probability" follows "weak convergence," but not vice versa.

functions $f_n(u) \to f(u)$, $f(0) = 1$ must hold; since $f(u)$ is continuous, one can choose a small positive number η such that

$$\left| \frac{1}{2\eta} \int_{-\eta}^{\eta} f(u)\, du \right| > 1 - \frac{\epsilon}{2} > \delta + \frac{\epsilon}{2}. \tag{79}$$

We shall now construct a contradiction to (79).

Take $X > 4/\eta\epsilon$ and consider $\delta_\nu = P_{n_\nu}(X) - P_{n_\nu}(-X)$; since the δ_ν converge toward δ we can choose an N so large that for $\nu > N$

$$\delta_\nu = P_{n_\nu}(X) - P_{n_\nu}(-X) < \delta + \frac{\epsilon}{4}; \tag{80}$$

also, by the definition of $f_{n_\nu}(u)$:

$$\int_{-\eta}^{\eta} f_{n_\nu}(u)\, du = \int_{-\eta}^{\eta} \left[\int e^{iux}\, dP_{n_\nu}(x) \right] du = \int \left[\int_{-\eta}^{\eta} e^{iux}\, du \right] dP_{n_\nu}(x), \tag{81}$$

where the order of integration has been reversed [as in (65)]. We wish to estimate the last integral. First, for all x, since $| e^{iu} | = 1$:

$$\left| \int_{-\eta}^{\eta} e^{iux}\, du \right| \leqslant 2\eta\,, \tag{82}$$

but on the other hand, if $x > X$:

$$\int_{-\eta}^{\eta} e^{iux}\, du = \frac{2}{x} \sin \eta x \leqslant \frac{2}{x} < \frac{2}{X}. \tag{83}$$

Applying now (82) for $| x | \leqslant X$ and (83) for $| x | > X$, we obtain from (81)

$$\begin{aligned} \left| \int_{-\eta}^{\eta} f_{n_\nu}(u)\, du \right| &\leqslant \left| \int_{|x| \leqslant X} \left[\int_{-\eta}^{\eta} e^{iux}\, du \right] dP_{n_\nu}(x) \right| \\ &\quad + \left| \int_{|x| > X} \left[\int_{-\eta}^{\eta} e^{iux}\, du \right] dP_{n_\nu}(x) \right| < 2\eta\delta_\nu + \frac{2}{X}. \end{aligned} \tag{81'}$$

Since from (80) $\delta_\nu < \delta + (\epsilon/4)$ and since X was to satisfy $X\eta > 4/\epsilon$, we have

$$\left| \frac{1}{2\eta} \int_{-\eta}^{\eta} f_{n_\nu}(u)\, du \right| < \delta + \frac{\epsilon}{2}, \tag{81''}$$

and in the limit

$$\frac{1}{2\eta} \int_{-\eta}^{\eta} f(u)\, du \,\Big| \leqslant \delta + \frac{\epsilon}{2}, \tag{84}$$

in contradiction to (79).

Hence, we have proved that the subsequence $\{P_{n_\nu}(x)\}$ tends toward *a distribution* $P(x)$ in every continuity point of $P(x)$. Then, by theorem (1) (which we have already proved) the limit $f(u)$ of the characteristic functions must be identical with the characteristic function of $P(x)$. By the uniqueness theorem, we know that the distribution is uniquely determined by $f(u)$ and therefore, it follows that *every* convergent subsequence of $\{P_n(x)\}$ tends to the limit $P(x)$. Hence $\{P_n(x)\}$ converges to $P(x)$ in every continuity point of $P(x)$. Theorem (2) is thus proved.

8.4. *A comment on Helly's theorem B.* We return once more to our proof in Section 8.3 of the main proposition (2). A main part of our proof of theorem (2) consisted in showing that *under the conditions of this theorem* the non-decreasing limit function $P(x)$ of Helly's theorem B is a distribution function. In general, this need not be so. Consider a sequence of densities $p_n(x)$, each zero outside the interval $(-n, +n)$ and equal to the constant $1/2n$ inside this interval. The corresponding $P_n(x)$ are equal to 0 for $x \leqslant -n$, to $P_n(x) = \frac{1}{2} + (x/2n)$ for $-n \leqslant x \leqslant n$, and $P_n(x) = 1$ for $x \geqslant n$. But $P(x) = \lim_{n\to\infty} P_n(x) = \frac{1}{2}$ for all x and $P(-\infty) = P(+\infty) = \frac{1}{2}$. Hence $P(x)$ is not a distribution function. On the other hand, the characteristic function of $P_n(x)$ is

$$f_n(u) = \frac{1}{2n}\int_{-n}^{n} e^{iux}\,dx = \frac{\sin nu}{nu},$$

and, as $n \to \infty$, this $f_n(u)$ tends to 1 for $u = 0$ and to 0 for all other u-values; hence $f(u)$ is not continuous and is not a characteristic function.

Problem 19. Prove that

$$f(u) = \frac{1}{k}\frac{\sin(ku/2)}{\sin(u/2)}\,e^{\frac{1}{2}i(k+1)u}$$

is the c.f. of the uniform distribution over the points $x = 1, 2, \ldots, k$.

8.5. *Applications.* We consider two applications. The first concerns a proof of Khintchine's theorem (see p. 234). Denote by $f(u)$ the c.f. of the distribution $P(x)$. Then the c.f. of $z = (1/n)(x_1 + \cdots + x_n)$ equals $[f(u/n)]^n$. From the expansion (46) we have

$$f(u) = 1 + imu + o(u),$$

and

$$\left(f\left(\frac{u}{n}\right)\right)^n = \left(1 + \frac{imu}{n} + o\left(\frac{1}{n}\right)\right)^n.$$

As $n \to \infty$ the right-hand side of this equation tends, for any fixed u toward e^{imu}. Hence by the continuity theorem the random variable z converges in probability to its mean value m.

As a second application we consider the proof of the *convergence of the binomial toward a normal distribution* which we presumed in Chapter IV. We seek the limit of $Q_n(x)$ as $n \to \infty$ while p, q remain fixed. Both mean value $a_n = nq$ and variance $s_n{}^2 = npq$ become infinite; hence, the number of successes, x, would have infinite mean and infinite standard deviation and no limit distribution could be found. Hence, we consider the standardized variable

$$y = \frac{x - a}{s} = \frac{x - nq}{\sqrt{npq}} \tag{85}$$

(also called the *reduced* variable) which has mean value zero and variance 1. From Problem 15 we take the characteristic function of y as

$$f_n(u) = (pe^{-iu\sqrt{q/np}} + qe^{iu\sqrt{p/nq}})^n. \tag{86}$$

We wish to expand the expression in parentheses in a Taylor series in u. For real z we have

$$e^{iz} = 1 + iz - \frac{z^2}{2} + \frac{(iz)^3}{3!} e^{\theta iz},$$

where $0 \leqslant \theta \leqslant 1$; we apply this to the two terms in parentheses and obtain easily

$$pe^{-iu\sqrt{q/np}} + qe^{iu\sqrt{p/nq}} = 1 - \frac{u^2}{2n}(1 + R_n),$$

where $R_n = O(1/\sqrt{n})$. Hence

$$f_n(u) = \left[1 - \frac{u^2}{2n}(1 + R_n)\right]^n .$$

From $\lim_{n\to\infty} [1 + (z/n)]^n \to e^z$, one can see that $\lim_{n\to\infty} f_n(u) = e^{-u^2/2}$ and, by example (1), Section 5.2, $e^{-u^2/2}$ is the characteristic function of $\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x} e^{-t^2/2}\,dt$. Therefore, on account of the continuity theorem (2) of Section 8.1, the c.d. function of the binomial distribution for the standardized variable $(x - nq)/\sqrt{npq} = y$ converges to $\Phi(y)$. Or, in a formula, with $q_n(x) = \binom{n}{x}q^x p^{n-x}$, $x = 0, 1, \ldots, n$ and $Q_n(x) = \sum_{\nu\leqslant x} \binom{n}{\nu} q^\nu p^{n-\nu}$, we obtain the Laplace-de Moivre result:

$$\lim_{n\to\infty} Q_n(nq + y\sqrt{npq}) = \lim_{n\to\infty} \Pr\left\{\frac{x - nq}{\sqrt{npq}} \leqslant y\right\} = \Phi(y). \tag{87}$$

or

$$\lim_{n\to\infty} \Pr\left\{a \leqslant \frac{x - nq}{\sqrt{nqp}} \leqslant b\right\} = \Phi(b) - \Phi(a). \tag{87'}$$

In Chapter VI we shall reach this result again as a particular case of a very general theorem. We shall also see that for $q_n(x)$ itself, a limit formula exists, namely,

$$\lim_{n\to\infty} \sqrt{npq}\,q_n(nq + y\sqrt{npq}) = \phi(y). \tag{88}$$

Problem 20. Two chance variables are each subject to the uniform distribution $p(x) = 1$ for $-\frac{1}{2} \leqslant x \leqslant \frac{1}{2}$. Find the distribution of their sum.

Problem 21. Prove by applying the recursion formula (24) that the sum of n normally distributed chance variables is also normally distributed.