CHAPTER IV

EXAMPLES OF COMBINED OPERATIONS

A. Uniform Distributions (Sections 1 and 2)

1. Uniform Arithmetical Distribution

1.1. *Definition. Examples.* We speak of a *uniform distribution* if, in the arithmetical case, each point of the label set carries the same probability value, and in the geometrical case, if the probability density is constant for all points x of the label set. In this introductory section, we will discuss the simple operations introduced in Chapter I as they apply to collectives with uniform arithmetical distributions.

Consider a distribution with the label set S consisting of the labels $x_1, x_2, ..., x_k$. All $p(x_i)$ being equal, it follows that $p(x_i) = 1/k$. Let $x_{i_1}, x_{i_2}, ..., x_{i_s}$ be a subset A of S. The probability $\Pr\{x \in A\}$, that is, the probability of the "event" that the chance variable x falls on one of the points $x_{i_1}, x_{i_2}, ..., x_{i_s}$, is given by the mixing rule as $p(x_{i_1}) + p(x_{i_2}) + \cdots + p(x_{i_s})$. Evidently this sum is s/k, a result which is usually interpreted in the following way.

Since all p's are equal, we may say that there are altogether k *equally possible* or *equally likely* outcomes in the original collective. Hence, among the equally likely outcomes there are s cases in which the event $x \in A$ occurs. We may call them the s "favorable" cases. The $\Pr\{x \in A\}$, which equals s/k, is then seen to be the *quotient of the number of favorable cases divided by the number of all possible cases.*

As an illustration, we refer to the game of roulette. Along the circumference of a wheel, 37 shallow grooves are equally distributed and colored alternately red and black, while one of the compartments is left white. The numbers 1, 2, ..., 36 are assigned in some way to the colored grooves, the white one receiving the number zero. The game consists of betting on the place at which a little ball that spins inside the rim of the wheel will come to rest. The mechanical conditions are assumed to be such that the ball may, with equal probability, come to rest at any one of the 37 compartments.

There are many ways to "make one's game" at the roulette table. One may bet that the ball will come to rest on a compartment that has an even number, or that it will come to rest on a "red compartment," or on one whose positive number is smaller than or equal to 18. In these three games, the number of "favorable" cases is 18, while the number of equally likely cases is 37. The probability of winning such a game is, therefore, 18/37. Suppose that T is a player's stake in a game of that type; if he wins, the rule of the game is that the reward R equals twice the stake, i.e., $R = 2T$. The expectation of this reward is $E[R] = 2T(18/37) + 0 \cdot (19/37) = (36/37)\,T$, that is, slightly smaller than his stake. Of course, the bank's business is built upon that difference.

Let us now apply the operation of partitioning to a collective with uniform distribution. Using the same notation as in Chapter I, we ask now for the conditional probability $P_A(B)$ for x to fall on one of the points $x_{j_1}, x_{j_2}, \ldots, x_{j_t}$ of a subset B of A, when it is already known that $x \in A$. We know that $P_A(B) = \Pr\{x \in B\}/\Pr\{x \in A\} = (t/k) \div (s/k) = t/s$. Again, we may interpret this result as a quotient of the number of favorable cases over the number of possible cases: the event $x \in A$ can be realized in s equally likely ways; among them, there are t "favorable" ones, i.e., t cases that realize the event $x \in B$. For example: if it is already known that the number appearing in a certain turn at the roulette table is a multiple of 5, and not zero, the probability that this number is either 10 or 15 is 2/7.

Finally, we turn to the operation of combination. Suppose we have two independent and combinable collectives K', K'', with the label sets $(x_1, x_2, \ldots, x_k)$ and $(y_1, y_2, \ldots, y_l)$, respectively. We assume that the respective distributions are $p'(x_i) = 1/k$ and $p''(y_j) = 1/l$, and consider the collective obtained by combination of K' and K''. We know that in the new collective the probability for any particular pair of label values (x_i, y_j) equals $p'(x_i) \cdot p''(y_j) = 1/kl$. The collective obtained by combination of two collectives with uniform distribution thus again has a uniform distribution. If we ask for $\Pr\{(x_i, y_j) \in A\}$ where A is now a certain subset containing s of the kl label values, this probability, according to the mixing rule, will be s/kl, again the quotient of favorable over possible cases.

As an illustration, take the probability of throwing a certain sum, r, with two true dice. There are 36 equally likely cases, (1, 1; 1, 2; ...; 1, 6; 2, 1; 2, 2; ...; 6, 6), among which $s = f(r)$ are favorable, $f(r)$ being the number of pairs with the sum r. Obviously, $f(r) = r - 1$ for $r \leqslant 7$ and $f(r) = 12 - (r - 1)$ for $r \geqslant 7$. Thus, for example, the probability of the sum 5 is $(5 - 1)/36 = \frac{1}{9}$.

1.2. *"Equally likely" cases. Problems.* In the language of the classical theory, as developed in the 17th century and presented as a complete system by Laplace (1812), each problem starts with the establishment of k equally likely cases, to which the probability $1/k$ is assigned. It is sometimes said that if we have no reason for the contrary assumption, we *have* to assume that all possible cases are equally likely (principle of indifference or of insufficient reason). Then the problem of finding the probability p of a certain "event" requires only finding the number s of those cases that are "favorable to the event," and then p equals s/k. This quotient s/k, favorable to possible cases, is given as the definition of the probability of an event. Since the correct formulation of this definition requires the use of the words "equally likely," it is a vicious circle because equally likely can only mean having the same probability. One can say that Laplace's definition reduces the general probability concept to a simple particular case, which remains undefined. But apart from this matter of logic Laplace's definition is only applicable in a very restricted field. No meaning can be given to the equally likely cases in defining the probability of throwing 6 with a biased die, or in the definition of probability in an insurance problem, etc.

We will now discuss two classical problems both posed by the Chevalier de Méré, a reputed gamester, that were among the earliest ones solved by the mathematicians of the 17th century (Pascal, Fermat).[1] Both problems arise from gambling situations and are typical of the sort of questions that attracted interest at the initial stage of probability calculus.

The first question is concerned with the fair division of a stake in an interrupted game. Suppose the two participants, A and B, in a game (not necessarily a game of chance) have agreed that whoever is the first to win a given number of turns gets the stake. The game, however, is discontinued at a moment when neither has won n turns. The question is how the stake should be divided. Suppose $n = 5$ and the players stop when A has won 4 turns and B 3 turns. According to the rules of the game, it would come to an end within the next two turns. To obtain the stake, B would have to win both of them, A only one of them. One might, therefore think that the stake should be divided in the ratio 2 : 1 between A and B; on the other hand, a partition in the ratio 4 : 3 (ratio of turns already won by A and B) might seem justifiable.

[1] These problems which were proposed to Pascal by the Chevalier are discussed in the correspondence of Fermat and Pascal. See I. Todhunter, *A History of the Mathematical Theory of Probability from the time of Pascal to that of Laplace.* 1st ed., Cambridge, 1865. Reprint: New York, 1931.

To obtain a solution that is correct in the sense of probability calculus, we have to find the probability that A or B will win the game. We must formulate this question in terms of collectives and their distributions and we do this in a way which makes clear the simple operations involved. We assume that A and B have the probabilities p and q, respectively, to win a single turn ($p + q = 1$). The label value assigned to each turn shall be a or b according to whether A or B wins this turn. The game will be finished after two further turns: either B wins them both, and then he has made it; or A wins at least once, then he has it. Hence, the collective on which the solution of our problem depends consists of successive pairs of single turns. As explained at the beginning of Section 8, Chapter I, this collective can be derived from the original collective as a combination of two collectives obtained by related place selections. The possible label values in this (two-dimensional) collective are a, a; a, b; b, a; b, b and their probabilities are p^2, pq, qp, q^2. Now the mixing rule must be applied, fusing together those pairs where A wins at least one turn. These are a, a; a, b; and b, a with the probability

$$p^2 + pq + qp = p(p + 2q). \tag{1}$$

This quantity is the probability for A to win the game. The only case in which B wins is that in which b, b happens, with the probability q^2. The sum of the two probabilities $p(p + 2q) + q^2$ equals $(p + q)^2 = 1$ as required.

If $p = q = \frac{1}{2}$ (uniform distribution in the original collective), the respective probabilities are $\frac{3}{4}$ and $\frac{1}{4}$ and the stake should be divided accordingly in the ratio 3 : 1, that is, neither in the ratio 2 : 1 nor 4 : 3.

Next, we consider the other problem of Chevalier de Méré. It is concerned with the advisability of betting on the appearance of "at least one 'six' in 4 casts of a correct die" (or in the simultaneous throwing of 4 correct dice) as compared with the appearance of "at least one 'double six' in 24 casts of a pair of correct dice." It was thought that the odds of the two games should be equal since the number of possible results with a pair of dice is 6 times that of a single die and 6 is also the ratio of the number of casts in the two games. In practice, however, it turned out that betting on "six" in the first game was advantageous, while betting on "double-six" in the second game was unfavorable. We start with the first question.

In our investigation we need not assume an unbiased die; instead, we assign in the collective K_0 the probability p_i to the face with i spots. Let us first reduce the collective K_0 to an *alternative*, i.e., a collective with two labels only, by mixing the labels 1, 2, ..., 5 into a single label, which we call 0 (non-six), and by denoting the label 6 by 1. Obviously in this collective K with labels 0, 1 we have $p(0) = p_1 + p_2 + \cdots + p_5 = p$ and $p(1) = p_6 = q$, with $p + q = 1$. The collective K' on which the problem depends is then obtained from K by the combination of four collectives, each obtained by means of a place selection (see again the beginning of Section 8 Chapter I). Each element of K' consists of four successive elements of K. The labels in K' are four-digit combinations of 0's and 1's. We want to find the probability of the combinations (by a mixing operation) that contain at least once the digit 1. Since the probability of the combination which consists only of 0's is p^4, the complementary probability $1 - p^4$ supplies the answer to our problem. This can be written as $1 - (1 - p_6)^4$, which in the case of an unbiased die becomes $1 - (\frac{5}{6})^4 = 0.516$.

The same way of reasoning gives the probability of throwing at least one "double-six" in 24 casts of a pair of dice. We have to replace the number 4 by 24 and the probability p_6 by the probability $p_{6,6}$ of a double six. The result is $1 - (1 - p_{6,6})^{24}$, which, for a pair of correct dice becomes $1 - (35/36)^{24} = 0.491$. Thus, the probability value is a little smaller than $\frac{1}{2}$ while the first was slightly above $\frac{1}{2}$. It is remarkable that the rather small deviations of these results from the value 0.500 are still large enough to be tangible in gambling.

In closing this section, we want to illustrate by a well-known example that the task of enumerating the "equally likely" cases is by no means always unequivocal. Let us reconsider *Bertrand's box problem* (Chapter I, Problem 15), assuming this time that the probabilities of selecting any one of the three types of boxes are equal ($=\frac{1}{3}$) and that the probabilities for opening the drawers of a selected box are also equal ($=\frac{1}{2}$). We ask now: if a box has been selected and one of its drawers opened and the coin inspected, what is the probability of finding in the other drawer of this box a coin of different metal? The obvious answer is $\frac{1}{3}$, since this can only happen when the box of type 3 has been selected, in which case it must happen.

However, the following way of reasoning has also been suggested. Suppose a gold coin had been found in the opened drawer of the selected box. The probability for this event is $\frac{1}{2}$, since there are altogether 3 gold coins and 3 silver coins. The other drawer in this box may hold gold or silver, according to whether the selected box was type 1 or type 3, respectively. Since these two cases are equally likely, the probability of finding silver in the other drawer is $\frac{1}{2}$. Hence, the probability for the compound event, finding gold in first and silver in second drawer, is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Since analogous reasoning applies to the event "finding silver in first and gold in second drawer," the probability of opening a drawer and finding a coin of different metal in the other drawer of the same box is $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

In the first argument, the assumption used is that the selection of any of the three boxes is equally likely, hence each is equal to $\frac{1}{3}$; since the probabilities for each drawer of the selected box are also equal, the probabilities for all six drawers are equal, hence each is equal to $\frac{1}{6}$. In the second argument it is assumed that if one drawer is opened and inspected the two possible contents of the other drawer represent equally likely cases. This, however, is not consistent with the assumption of $\frac{1}{6}$ as the probability for each drawer; in fact, if we already know that the drawer which was opened contained a gold coin, then this drawer could be either the gold drawer of the third box, or one or the other drawer of the first box and since these are 3 choices, the probability for the first of these events is $\frac{1}{3}$ and not $\frac{1}{2}$. This simple example shows the danger of an indiscriminate use of "equally likely" cases. The unequivocal way to deal with the box problem is to use the partitioning rule as indicated in Problem 15 of Chapter I; then the result is clearly seen to be dependent on the assumptions; in our case, the probability that the gold coin found is from the gold drawer of box III is $p_3q/(p_1 + p_3q)$ which equals $\frac{1}{3}$, if $p_1 = p_2 = p_3 = \frac{1}{3}$, $q = \frac{1}{2}$.

Problem 1. What is the probability that, in casting two unbiased dice, the difference between the two spot-values has the value $d = 0, \pm 1, \pm 2, ..., \pm 5$?

Problem 2. An unbiased coin is thrown n times. Compute the probability that tail and head appear alternately.

Problem 3. Prove that in casting three correct dice simultaneously the probability is $\frac{1}{2}$ that the sum of the three spot-values is greater than 10.

Problem 4. In an election the candidates A and B received a and b votes, respectively, with $a > b$. In counting the votes, the ballots are drawn successively at random from the ballot box. Assume that each of the possible $(a + b)!/a!b!$ permutations (see p. 168) has the same probability. Prove that the probability that throughout the counting, A is permanently ahead of B equals $(a - b)/(a + b)$ (Bertrand).[2]

2. Uniform Density. Needle Problem

2.1. *Equally likely cases.* In a certain way, the statements developed for uniform arithmetical distributions can be extended to the case of a constant probability density. Assume that a chance variable x can assume all values between a and b and that the density p within this interval is

[2] See discussion and historical remarks in Feller [7b], p. 66 ff.

constant and equals zero outside the interval. Then, since $\int_a^b p\,dx$ must equal one, p has the value $1/(b-a)$. If (x_1, x_2) is some partial interval of (a, b), the probability that x falls into (x_1, x_2) equals

$$\int_{x_1}^{x_2} p\,dx = p\int_{x_1}^{x_2} dx = \frac{x_2 - x_1}{b - a}. \tag{2}$$

The denominator can be considered as a measure of all "possible cases" and the numerator as the analogous measure of all "favorable cases," namely, favorable to the event that x falls in the interval (x_1, x_2). Thus, one can again say that the probability of the event under consideration is the quotient of the "number" of favorable cases to that of all possible cases, provided these are all equally likely.

While this is undoubtedly correct in the case of Eq. (2), one is led here to new difficulties if one tries to base the whole of probability theory upon the concept of equally likely cases. The following is a simple example of these difficulties.

Assume that a glass contains a mixture of wine and water. It may be known that the ratio x of wine to water lies between 0.5 and 1.0. If all intermediate cases are "equally likely," there is according to (2) a chance of 0.50 of x falling in the interval 0.5 to 0.75 (and the same chance for the interval 0.75 to 1.0). In this problem, the ratio y of water to wine lies obviously in the interval 1 to 2. Since all intermediate cases are assumed to be "equally likely," it follows that the chance is 0.50 to find y in the interval 1.5 to 2 (and the same for the interval 1 to 1.5). Now, since $xy = 1$, the interval 1.5 to 2 for y corresponds to the interval 0.5 to 0.666... of x. Thus, we have arrived at an inconsistency: first, we found a 50% chance for x falling in the interval 0.5 to 0.75 and then, apparently under the same conditions, a 50% chance for x falling in the smaller interval 0.5 to 0.666 ... . The obvious explanation is that if a continuous chance variable is involved, the expression "equally likely cases" is equivocal. The distribution in the original collective must be given in a precise way; otherwise nothing can be computed.

Let $P(x)$ be the probability distribution function of x. If the probability density with respect to x, that is, dP/dx, is known to have a constant value, then the first result holds. On the other hand, if the density with respect to y, that is, dQ/dy, is constant, where $Q(y)$ is the probability distribution of y, the second result is correct. The two assumptions are incompatible since, from $xy = 1$ and on account of [1]

$$P(x) = \Pr\left\{\frac{\text{wine}}{\text{water}} \leqslant x\right\}, \qquad Q(y) = \Pr\left\{\frac{\text{water}}{\text{wine}} \leqslant y\right\} = \Pr\left\{\frac{\text{wine}}{\text{water}} \geqslant x\right\},$$

[1] As in the preceding chapter we use the notation Pr{ } or Prob{ } if the expression

we conclude $P(x) + Q(y) = 1$ and

$$-\frac{dQ}{dy}\frac{1}{x^2} + \frac{dP}{dx} = 0 \quad \text{or} \quad \frac{dQ}{dy} = \frac{dP}{dx}x^2. \tag{3}$$

2.2. *Buffon's needle problem.* An example that played a role in the history of probability calculus is known as *Buffon's needle problem.* A small needle of length b is thrown at random onto a horizontal board on which parallel lines with a spacing $a > b$ are drawn. What is the probability that the needle crosses one of the straight lines (see Fig. 6) ? Again, all possible positions of the needle are considered "equally likely cases." Any position of the needle can be determined by the two abscissas x_1 and x_2 of its end points. (The ordinates are irrelevant to the question of

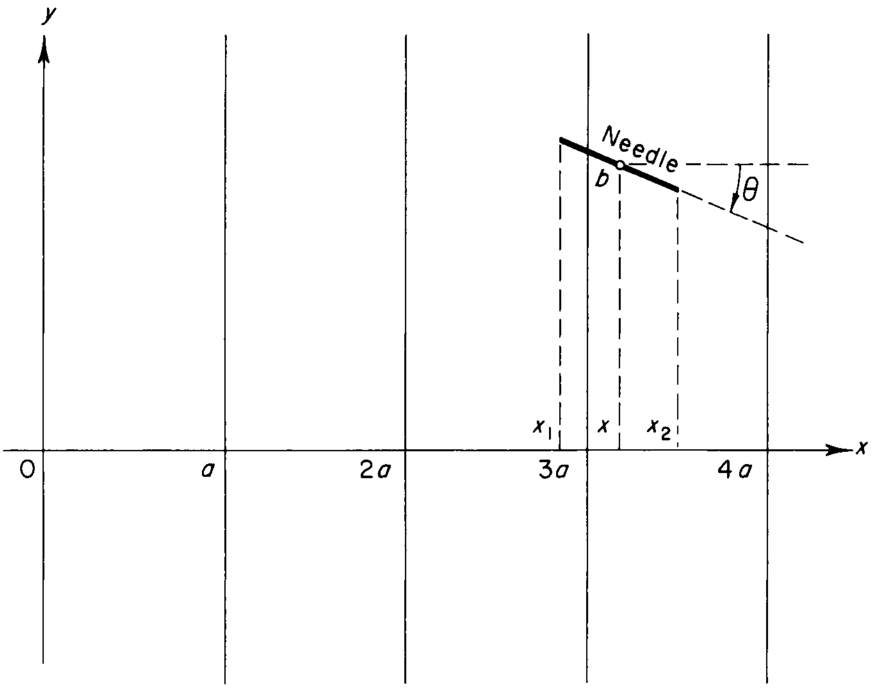

FIG. 6. Needle problem.

between the parentheses stands for a sentence; in the present case $\Pr\left\{\frac{\text{wine}}{\text{water}} \leqslant x\right\}$ stands for "probability of the proportion wine/water being less than or equal to x." We write $P(A)$, $P(B)$, if the argument is A or B. If we wish to explain the meaning of the set A, then Pr{ } would take the place of $P(A)$.

whether or not the needle crosses the straight line.) If we admit only those cases in which the center of the needle lies within the board extending from $x = 0$ to $x = na$ (that is, including $n + 1$ lines), the range of possible x_1, x_2-values is restricted by the inequalities

$$0 \leqslant \frac{x_1 + x_2}{2} \leqslant na, \qquad 0 \leqslant x_2 - x_1 \leqslant b.$$

In an x_1, x_2-plane, this region is represented by a rectangle whose boundaries have the equations

$$x_1 + x_2 = 0, \qquad x_1 + x_2 = 2na, \qquad x_2 - x_1 = 0, \qquad x_2 - x_1 = b.$$

This is illustrated in Fig. 7, where n is assumed to be 4. The measure of all possible cases is the area of this rectangle and, hence, equals nab.

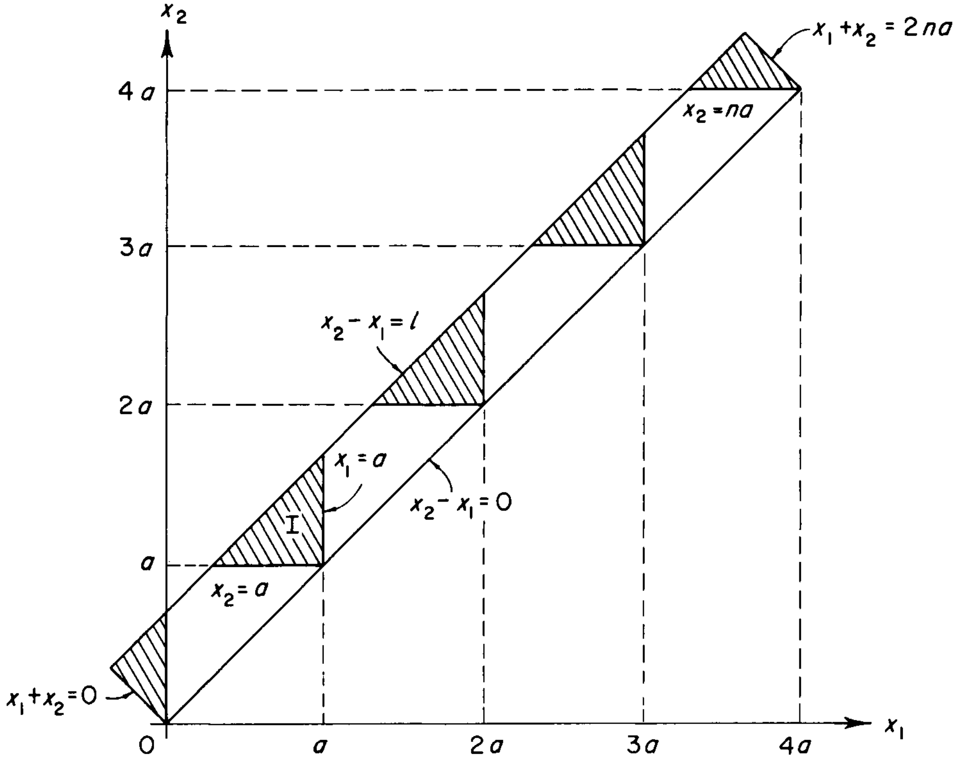

FIG. 7. Needle problem. Regions in x_1, x_2-plane.

A needle crosses the line $x = a$ if, and only if, $x_1 < a$ and $x_2 > a$. The two lines $x_1 = a$ and $x_2 = a$ in Fig. 7 cut out the shaded triangle I, whose area is $b^2/2$. All cases favorable to a crossing are represented by the shaded regions, which cover a certain part of the rectangle of area nab. The total shaded area is $nb^2/2$. Therefore, the chance for a needle to cross one of the n parallel lines on the board would be, according to (2),

$$\frac{nb^2/2}{nab} = \frac{b}{2a}, \tag{4}$$

that is, proportional to the length b of the needle and inversely proportional to the distance a.

We can, however, determine a needle's position in a different way by two other parameters. For example, we may use the abscissa x of the center of the needle and the angle θ between the x-axis and the needle. The range of these variables is determined by

$$0 \leqslant x \leqslant na, \qquad -\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2}. \tag{5}$$

In a plane with the coordinates x, θ, this range is represented by the rectangle shown in Fig. 8 whose area is $na\pi$. The needle will cross the line $x = a$ if, and only if,

$$-\frac{b}{2}\cos\theta < x - a < \frac{b}{2}\cos\theta. \tag{6}$$

The two curves corresponding to $x = a \pm (b/2)\cos\theta$ are plotted in Fig. 8; the region I cut out by these lines from the rectangle has the form of a convex lens of area

$$2\int_{-\pi/2}^{\pi/2} \frac{b}{2}\cos\theta\, d\theta = 2b. \tag{7}$$

The total shaded area representing all favorable cases consists of $n - 1$ such lenses and two half lenses as shown in Fig. 8. The magnitude of the total area is $2nb$. Thus, the chance for a needle to cross one of the n lines would amount to

$$\frac{2nb}{na\pi} = \frac{2}{\pi}\frac{b}{a} = 0.637\frac{b}{a}, \tag{8}$$

which is a little larger than the former value $0.5\ b/a$.[1]

The explanation of this apparent contradiction is included in what was said above about the use of the two ratios x and y in the problem of the wine-water mixture. The assumption that the probability density

[1] Buffon used the experiment to find an approximate value of π : a first instance of a Monte Carlo method.

with respect to x_1, x_2 is constant is incompatible with the hypothesis that the density with respect to x, θ is constant. Which one of these or many other assumptions should be made is a question of fact and depends on how the needles are thrown. It is not a problem of probability calculus to decide which distribution prevails in the original collective. The conclusion to which the theory leads is that the chance of crossing is given by the integrals

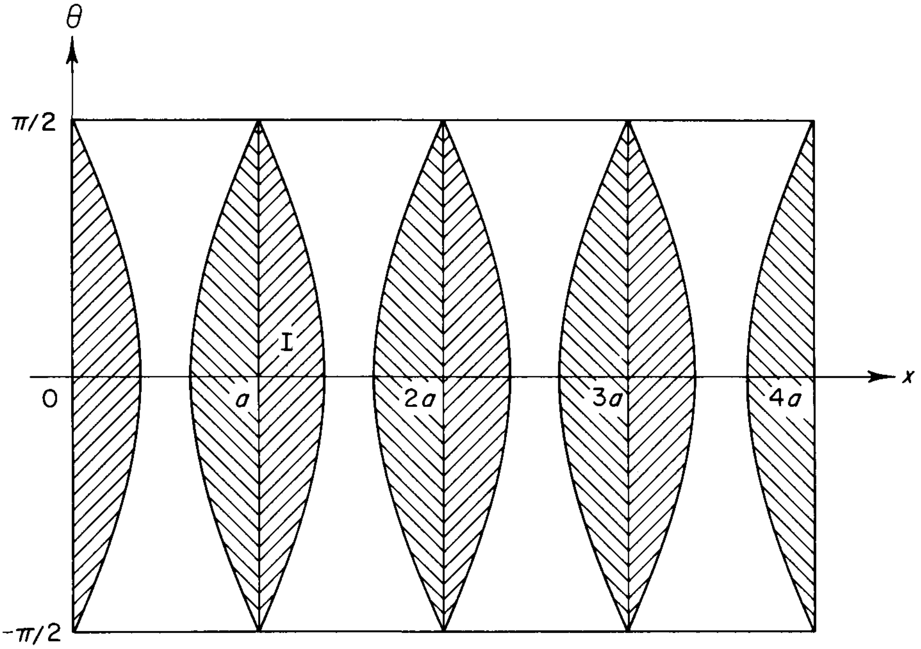

FIG. 8. Needle problem. Regions in x, θ-plane.

$$\int_{(S')} p'(x_1, x_2)\, dx_1\, dx_2 = \int_{(S'')} p''(x, \theta)\, dx\, d\theta, \tag{9}$$

where p' and p'' are the densities with respect to x_1, x_2 and to x, θ, respectively, and S', S'' the shaded regions in Figs. 7 and 8. Neither of the quantities p' and p'' need be constant. Large-scale experiments carried out under specified conditions have led to a frequency value closer to the result based on the second assumption. This does not imply that for a different experimental arrangement the first assumption would not be more adequate.

A certain branch of geometry known as "integral geometry" is concerned with the problem of finding an intrinsic measure for certain geometrical objects, like straight lines, triangles, convex regions, etc. (W. Blascke, *Vorlesungen über Integralgeometrie*. Erstes Heft, Leipzig, Berlin, 1935.) A measure of a set of straight lines in a plane, say, is called intrinsic if its value does not change when all lines are subject to the same displacement. If only translations in the x-direction and rotations are admitted, it can immediately be seen that $\iint dx\, d\theta$ is an intrinsic measure, where x is the abscissa of one point of a line and θ the angle of the line with a fixed direction, while $\iint dx_1\, dx_2$, with x_1 and x_2 the coordinates of two points, would change its value on rotation. In this way, as long as geometrical objects are concerned, certain sets of parameters are singled out; in our case, this could favor x, θ rather than x_1, x_2. But although the above-mentioned experiments also seem to favor x, θ, there is, nevertheless, no stringent reason for assuming that the probability density must be constant with respect to the intrinsic coordinates. Besides, in problems of a nongeometrical character like wine-water mixture no such indication is supplied for preferring either x or y. It remains invariably true that in any problem of probability calculus, the distribution in the original collective must be given. Only from such given data can results be derived.

Problem 5. A circle of radius 1 is plotted on a sheet of paper and a straight line intersecting the circle is drawn at random. What is the probability that the chord length is smaller than the radius? Assume the density constant (a) with respect to the distance of the straight line from the center and to its angle; (b) with respect to the positions (the polar angles) of the two intersection points.

Problem 6. Three points are chosen at random on the periphery of a circle. For each point, the probability of falling in an angular interval of length ϕ is $\phi/2\pi$. What is the probability that the three points lie on the same side of some diameter?

Problem 7. The interval 0, l is subdivided into three parts by the random choice of two points with abscissas x and y. Find the probability that the three partial lengths can be the sides of a triangle. The assumption is that the probability density is constant with respect to x and y.

B. Bernoulli Problem and Related Questions (Sections 3-6)

3. The Problem of Repeated Trials

3.1. *Problem and solution.* Consider a sequence K_0 of observations, each of which may have only one of two possible results. We might, for example, think of subsequent drawings, with replacement after each drawing, from an urn containing black and white balls, and shall denote the two possible label values by 0 and 1 (0 = black ball, 1 = white ball). Instead of speaking of the appearance of the labels 1 or 0 in each observation, we shall also use the terms occurrence or non-occurrence of a specified event (drawing of a white ball). We assume the existence of a certain probability q for that event, such that $p = 1 - q$ is the probability for non-occurrence of the event.[1]

Our problem is to determine the probability that, in a group of n of these trials, the event occurs x times ($x = 0, 1, ..., n$). We call this probability, which depends on p, q, and on the parameter n,

$$p_n(x) = \Pr\left\{\begin{matrix}\text{1 appearing } x \text{ times in a group of } n \text{ trials, each trial} \\ \text{having probability } q \text{ for 1 and } p = 1 - q \text{ for 0}\end{matrix}\right\}. \qquad (10)$$

The collective with the distribution $p_n(x)$ can be constructed in the following manner.

First, subdivide the original sequence K_0 into subsequences of length n. Then, form n auxiliary collectives K_1, K_2, ..., K_n by means of place selections, where the sequence K_i, $i = 1, 2, ..., n$ consists of the trials numbered $i, n + i, 2n + i, ...$. The probabilities for 0 and 1 are still p and q in each K_i according to the rule of selection. Next, we combine the independent collectives K_1, ..., K_n into one n-dimensional collective K', each element of K' thus accounting for n elements of K_0, such that the jth element of K' consists of the elements of K_0 numbered $n(j - 1) + 1$, $n(j - 1) + 2$, ..., nj, $j = 1, 2, ...$. The label set of K' thus consists of the 2^n possible combinations of 0's and 1's (the 2^n corners of an n-dimensional cube). The 2^n probabilities in K' are—by the multiplication rule—products $p^\alpha q^\beta$, where $\alpha + \beta = n$. Since there are only $n + 1$ different probability values assigned to 2^n possible labels, several labels must have the same probability value if $n \geqslant 2$.

Finally, we apply the following mixing. Out of the 2^n possible n-dimensional labels we fuse together those that correspond to exactly x successes, that is, contain the figure 1 exactly x times, regardless of its position. We call the new collective K, take x as the new label value, and $p_n(x)$ as the corresponding probability.

Generally, $p_n(x)$ for each value of x, is an aggregate of equal terms $p^{n-x}q^x$. The number of such terms equals the number of ways of placing

[1] Also the terms "success" and "failure" are used occasionally. In most of his publications, v. Mises uses p, q for the probabilities of 0 and 1, respectively: $p(0) = p$, $p(1) = q$, since q follows p as 1 follows 0. Since "1" stands for "event" or "success" one arrives at the above notation.

indistinguishable objects (one's) on n places. Since this number is[2] $\binom{n}{x} = \frac{n(n-1)\cdots(n-x+1)}{1\cdot 2\cdot 3\cdots x} = \frac{n!}{x!\,(n-x)!}$. The probability of x events in n trials, or the probability of obtaining the sum x in adding n consecutive labels 0 or 1 is therefore

$$p_n(x) = \binom{n}{x} p^{n-x}q^x, \qquad x = 0, 1, \ldots, n. \tag{11}$$

We note that $\binom{n}{x} = \binom{n}{n-x}$; $\binom{n}{0} = \binom{n}{n} = 1$ in accordance with $0! = 1$.[3] The discontinuous distribution $p_n(x)$, determined by (11), is called the *Bernoulli distribution* or *binomial distribution*, since it equals the xth term in the binomial expansion of $(p + q)^n$ (see Fig. 9).

3.2. *Multinomial distribution.* Before considering the distribution (11) more closely, we note that a very slight change of our considerations leads to the *multinomial distribution*. Assume a sequence of independent observations which may have k different results, for example, the successive drawing with replacement from an urn which contains balls of k different colors in the proportions $p_1 : p_2 : \cdots : p_k$, where $p_1 + p_2 + \cdots + p_k = 1$. Denote by $p_n(x_1, x_2, \ldots, x_k)$ the probability of obtaining in n drawings x_1 balls of the first color, x_2 of the second, ..., x_k of the last one, where $x_1 + x_2 + \cdots + x_k = n$. It is then easily seen [by induction, starting from (11)] that

$$p_n(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}, \tag{12}$$

for non-negative integers $x_1, x_2, \ldots, x_k$, with $x_1 + x_2 + \cdots + x_k = n$. Note that on account of this last equality, the distribution given in (12) is

[2] One arrives at this well-known result in the following way: first, we place one object; this can be done in $n_1 = n$ ways. There remain $n - 1$ places, hence there are $n(n-1)$ possibilities for two objects. But since the objects are identical the arrangements equal each other in pairs; hence $n_2 = n(n-1)/2$. For the third object there remain $n - 2$ places and we obtain all arrangements by combining n_2 possibilities for the first two objects with the $n - 2$ possibilities for the third one. In this way, however, any triple of places is counted three times since at each of the three places the "third" object may stand. Hence, $n_3 = (\frac{1}{3})n_2(n-2)$. In the same way $n_x = (1/x)\cdot n_{x-1}(n - x + 1)$ which gives the result.

[3] Logarithms of $n!$ and logarithms of the binomial coefficients $\binom{n}{x}$ are in A. Hald, *Statistical Tables and Formulas*, New York, 1952, Table XIII, $n = 1, \ldots, 1000$; Table XIV, $n = 2, \ldots, 100$, $x = 1, \ldots, 50$. See our small Table III.

a $(k - 1)$-dimensional distribution, just as for $k = 2$, we obtained the one-dimensional distribution (11). For reasons of symmetry we use the notation (12) if $k \geqslant 3$.

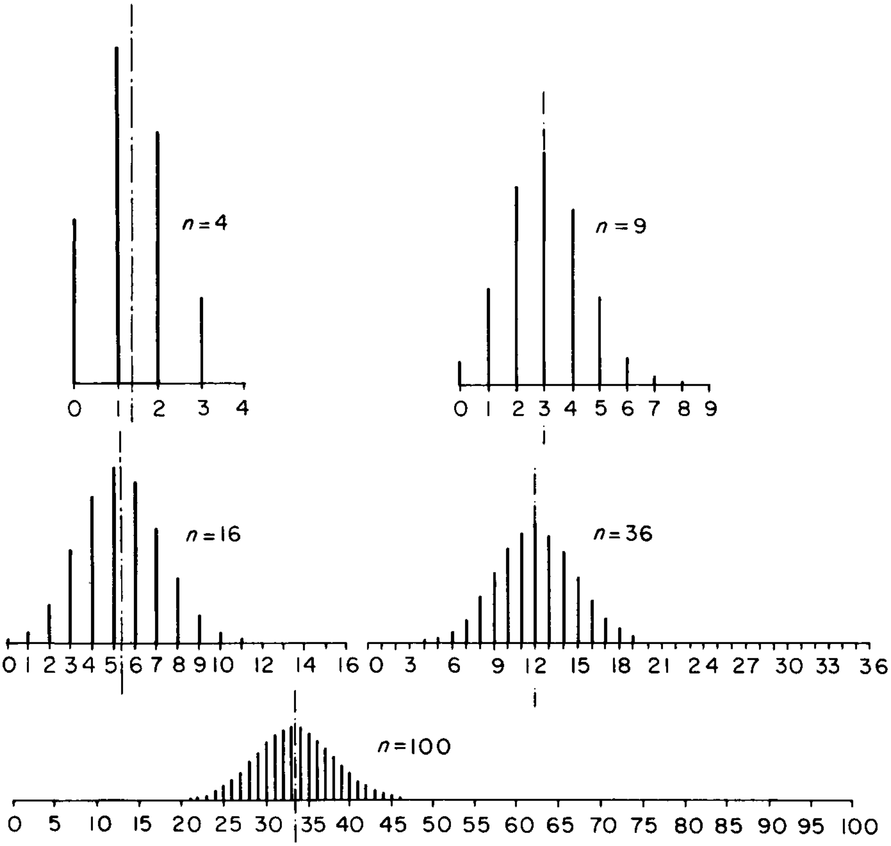

FIG. 9. Bernouilli distributions.

3.3. *Moments and mode of the binomial distribution.* We return to the study of the distribution (11) and, for that purpose, introduce the expression $(p + qt)^n$, where t denotes an independent variable.[4] The well-known binomial formula

$$(p + qt)^n = \binom{n}{0}p^n + \binom{n}{1}p^{n-1}qt + \cdots + \binom{n}{n}q^n t^n \tag{13}$$

[4] The method based on this expression will appear in Chapter V, in a general way.

may be written on account of (11) as

$$(p + qt)^n = p_n(0) + p_n(1)t + p_n(2)t^2 + \dots + p_n(n)t^n$$
$$= \sum_{x=0}^{n} p_n(x)\, t^x. \tag{13'}$$

In this equation, which is an identity in t, we may set $t = 1$ and obtain

$$p_n(0) + p_n(1) + \cdots + p_n(n) = (p + q)^n = 1, \tag{14}$$

which confirms that the sum of all $p_n(x)$ for a given n must equal 1.[5]

The function $(p + qt)^n$ which we call the moment generating function can be used to great advantage for the determination of the moments of the distribution $p_n(x)$. For that purpose, let us differentiate the identity (13′) with respect to t. The left-hand side yields $(d/dt)(p + qt)^n = n(p + qt)^{n-1} q$. Differentiating the right-hand side term by term, we obtain the relation

$$n(p + qt)^{n-1} q = p_n(1) + 2p_n(2)t + 3p_n(3)t^2 + \cdots + np_n(n)t^{n-1}. \tag{15}$$

On setting $t = 1$ again, the right-hand side becomes $\Sigma_{x=0}^{n} xp_n(x,)$ which is the mean value of the distribution $p_n(x)$ or the expectation of x. To the left, we have simply nq, since $p + q = 1$. Thus, we arrive at

$$\sum_{x=0}^{n} xp_n(x) = nq \qquad \text{or} \qquad a_n = E_n[x] = nq. \tag{16}$$

(The summation in $\Sigma_x\, xp_n(x)$ and similar expressions may be thought of as extending over all integral values of x since $p_n(x)$ is zero by definition for $x < 0$ and $x > n$.)

By repeating the process of differentiation, we obtain the value of the second moment and the variance of $p_n(x)$. In fact

$$\frac{d^2}{dt^2}[(p + qt)^n] = n(n - 1)\, q^2(p + qt)^{n-2}$$
$$= 2 \cdot 1p_n(2) + 3.2p_n(3)t + \cdots + n(n - 1)p_n(n)t^{n-2}. \tag{17}$$

If again 1 is substituted for t, the expression on the left-hand side

[5] The reader will also verify without difficulty that, in analogy to (13′) the probability $p_n(x_1, x_2, \dots, x_n)$ of (12) is the coefficient of $t_1^{x_1} t_2^{x_2} \dots t_k^{x_k}$ in the expansion of the polynomial $(p_1t_1 + p_2t_2 + \cdots + p_kt_k)^n$ in powers of the t_i, and that the sum of all possible values of (12) equals one.

becomes $n(n-1)q^2$, and that on the right-hand side of (17) is the sum of all terms $x(x-1)p_n(x)$. Thus

$$n(n-1)q^2 = \sum_x x(x-1)p_n(x) = \sum_x x^2 p_n(x) - \sum_x x p_n(x)$$
$$= \sum_x x^2 p_n(x) - a_n . \tag{18}$$

This gives for the second moment of $p_n(x)$

$$\sum x^2 p_n(x) = nq[(n-1)q+1] = nq(nq+p) = n^2q^2 + npq. \tag{19}$$

From the shift-of-origin rule, we find now the variance $s_n{}^2$ of the distribution $p_n(x)$:

$$s_n{}^2 = \sum_x x^2 p_n(x) - a_n{}^2 = npq. \tag{20}$$

This result is of great importance.

It can easily be seen that the higher derivatives of $(p+qt)^n$ lead to expressions for the higher moments of $p_n(x)$. In fact, the kth derivative expressed in the two ways as in (17) supplies (with $t = 1$)

$$n(n-1)(n-2)\cdots(n-k+1)\,q^k = \sum_x x(x-1)(x-2)\cdots(x-k+1)p_n(x). \tag{21}$$

The right-hand side is the kth factorial moment of $p_n(x)$; if the multiplication $x(x-1)(x-2)\ldots$ is carried out, the sum in (21) appears as a linear function of the zero moments[6] $m_n^{(1)}, m_n^{(2)}, \ldots, m_n^{(k)}$ of $p_n(x)$. Thus, if we write (21) for $k = 1, 2, \ldots, \mu$ we obtain μ linear equations for the μ unknowns $m_n^{(1)}, m_n^{(2)}, \ldots, m_n^{(\mu)}$ and from them the moments with respect to any other reference point, for instance the mean value, can be found. A shorter way of computing the moments of higher order will be presented in Chapter V.

Denote by $M_n^{(\mu)}$ the μth central moment (moment with respect to the mean value) of $p_n(x)$. It can be shown that the following symbolic formula holds:

$$M_n^{(\mu)} = (M_{n-1}^{(1)} + M_1^{(1)})^\mu ,$$

where the μth power of $M_1^{(1)}$ or of $M_{n-1}^{(1)}$ is to be replaced by $M_1^{(\mu)}$ or $M_{n-1}^{(\mu)}$.

For $\mu = 1$ this gives

$$M_n^{(1)} = M_{n-1}^{(1)} = \cdots = M_1^{(1)} = \sum_{x=0}^{1} (x - a_1)p_n(x) = 0;$$

[6] The notation for the moments differs from that in Chapter III, Section 4, since we need here the subscript n which corresponds to $p_n(x)$.

for $\mu = 2$ we have, since $M_1^{(1)} = 0$

$$M_n^{(2)} = M_{n-1}^{(2)} + pq = M_{n-2}^{(2)} + 2pq = \cdots = npq,$$

as in (20). For the third and fourth moments we find

$$M_n^{(3)} = M_{n-1}^{(3)} + qp^3 - pq^3 = M_{n-1}^{(3)} + pq(p - q) = \cdots = npq(p - q),$$

$$M_n^{(4)} = M_{n-1}^{(4)} + 6(n - 1)p^2q^2 + pq(1 - 3pq) = \cdots$$
$$= 3n(n - 1)p^2q^2 + npq(1 - 3pq) = 3n^2p^2q^2 + npq(1 - 6pq).$$

We may check the correctness of the last formula by computing the difference of the expressions for n and for $n - 1$. In the same way we find

$$M_n^{(5)} = npq(p - q)(10npq - 12pq + 1),$$

$$M_n^{(6)} = 15n^3p^3q^3 - 5n^2p^2q^2(26pq - 5) + npq(120p^2q^2 - 30pq + 1).$$

A general result is: *$M_n^{(2\mu)}$ and $M_n^{(2\mu+1)}$ are polynomials in n of degree μ.* This is certainly true for $\mu = 0$, since both $M_n^{(0)}$ and $M_n^{(1)}$ are independent of n. The general result follows by induction.[7]

We wish to discuss now the dependence of the function $p_n(x)$ on the number of successes x for a given number of trials n. For that purpose, consider the quotient of two successive values of $p_n(x)$:

$$\frac{p_n(x)}{p_n(x-1)} = \frac{n!(x-1)!(n-x+1)!}{x!(n-x)!n!}\,\frac{p^{n-x}}{p^{n-x+1}}\,\frac{q^x}{q^{x-1}} = \frac{n-x+1}{x}\,\frac{q}{p}. \tag{22}$$

The function $p_n(x)$ will increase (or decrease) with increasing x if $(n - x + 1)\,q$ is larger than (or smaller than) px. Or (using $p + q = 1$)

$$\begin{aligned} &p_n(x) \text{ increases,} \quad \text{if } x < (n+1)q, \\ &p_n(x) \text{ decreases,} \quad \text{if } x > (n+1)q. \end{aligned} \tag{23}$$

One sees that $p_n(x)$ will first increase and later decrease; therefore, a maximum value of $p_n(x)$ must exist. If x_m is an x-value for which $p_n(x)$ reaches its maximum, we shall call x_m the *most probable value* of x. If only one x_m exists, it must satisfy the two conditions

$$\frac{p_n(x_m)}{p_n(x_m - 1)} > 1, \qquad \frac{p_n(x_m + 1)}{p_n(x_m)} < 1, \tag{24}$$

which are equivalent, according to (22), to

$$nq - p < x_m < nq + q. \tag{24'}$$

The interval $(nq - p,\ nq + q)$ or $(a_n - p,\ a_n + q)$ has unit length and includes the point a_n. If its boundaries are not integers, x_m is uniquely

[7] v. Mises [21], p. 134.

determined by (24′) as the only integer in this interval; if a_n is an integer, $x_m = a_n$; if a_n is not an integer, then x_m is one of the two integral neighbors of a_n. If, however, $a_n - p$ (and therefore $a_n + q$) is an integer, then a_n is not an integer and two most probable values of x exist, with a_n between them. In fact, if $a_n + q$ is introduced for x in (22), we find the last fraction in (22) equal to one; that is, both integral neighbors of a_n have the same probability and this is evidently the greatest value $p_n(x)$ can take. In any event, *the mean value of x and the most probable value of x* (the latter being always an integer) *differ by less than unity.*

Problem 8. What is the probability of obtaining either one or two or three sixes in 10 throws of a correct die?

Problem 9. Let x_i $(i = 1, 2, ..., 6)$ be the number of results equal to i in n throws of a die, and p_i be the probability of throwing i spots with this die. Determine the expectation of the sum $x_1 + x_2 + x_3$.

Problem 10. Prove that the following holds for the multinomial distribution: if $i, j = 1, 2, ..., k$ and E denotes the expectation with respect to the multinomial distribution, then $a_j = E[x_j] = np_j$, $s_{jj} = E[(x_j - np_j)^2] = np_j(1 - p_j)$, $s_{ij} = E[(x_i - np_i)(x_j - np_j)] = -np_ip_j$, $i \neq j$.

Problem 11. Let p_i, $i = 0, 1, ..., k$ be the probability of obtaining the result i in a drawing from an urn, where $\Sigma_{i=0}^{k} p_i = 1$. This drawing is repeated n times with replacement. Denote by $q_n(x)$ the probability of obtaining the sum x in these n drawings.

(a) Prove that $(p_0 + p_1t + \cdots + p_kt^k)^n = \Sigma_{x=0}^{nk} q_n(x)\, t^x$.

(b) By the method used in this section find the mean value and variance of $q_n(x)$ for $k = 2$.

4. Bernoulli's Theorem

4.1. *Problem and solution.* We arrive at most significant conclusions if in $p_n(x)$ we slightly change the variable in question. So far, we have asked for the probability distribution of the number x of events in n trials. Let us now consider the ratio

$$z = \frac{x}{n} \tag{25}$$

which is the relative number (or the frequency) of the events (successes or "ones") in a group of n trials. Evidently, if $p_n'(z)$ is the probability of a z-value, one has

$$p_n'(z) = p_n(nz) = p_n(x). \tag{26}$$

Now, from formulas (16) and (20) for a_n and $s_n{}^2$ we derive

$$a_n' = E[z] = \sum_z z p_n'(z) = \sum_x \frac{x}{n} p_n(x) = \frac{a_n}{n} = q,$$

$$\begin{aligned} s_n'^2 = E[(z - a_n')^2] &= \sum_z (z - a_n')^2 p_n'(z) = \sum_x \frac{(x - a_n)^2}{n^2} p_n(x) \\ &= \frac{s_n{}^2}{n^2} = \frac{pq}{n}. \end{aligned} \tag{27}$$

Moreover, it is clear that the most probable value of z is x_m/n, x_m being the most probable value of x.

We infer from these results three statements:

(1) The expected value of the relative number z of successes in n trials is independent of n and equals the probability q of the event.

(2) The most probable value of z differs from its mean value q by less than $1/n$; that is, the two coincide in the limit of increasing n [see Eq. (24′)].

(3) The variance of z decreases with increasing n toward zero.

The significance of the last statement becomes more evident if we apply Tchebycheff's inequality. Let ϵ be any small positive quantity. Then

$$\Pr\{| x - q | \leqslant \epsilon\} > 1 - \frac{s_n'^2}{\epsilon^2} = 1 - \frac{pq}{n\epsilon^2}. \tag{28}$$

The last term tends to zero as n increases, whatever the value of the fixed number ϵ. Thus, we have proved the famous *Bernoulli Theorem* (1713):

However small ϵ is, as n increases, the probability approaches 1 *that the relative number $x/n = z$ of successes in n trials differs by less than ϵ from its expected value q.*

If, in this statement, we replace the word probability by its frequency-definition, we see that the statement relates to a large number of "experiments" where each experiment consists of a group of n single trials (each trial is an alternative with probability q for success). The Bernoulli theorem states that in the vast majority of these experiments, the observed relative number z of successes will be very close to q if n is correspondingly large. Let ϵ be given, and denote by $N_1(n)$ the number of these experiments out of N for which $| (x/n) - q | \leqslant \epsilon$. Then for any n, $\lim_{N\to\infty} N_1(n)/N = P_n$ is the $\Pr\{| (x/n) - q | \leqslant \epsilon\}$. Our statement is that $\lim_{n\to\infty} P_n = 1$.

4.2. *Discussion.* If, instead of the term "relative number of events," we use for z the expression "frequency," then a less careful formulation of the Bernoulli theorem is likely to lead to confusion. The quantity q was defined by us as the limit of the frequency of ones in the original sequence of zeros and ones. Now, the Bernoulli theorem states that for large n the frequency of ones is, with a probability nearly 1, in the neighborhood $q \pm \epsilon$. Have we come back to our point of departure? Or, must we modify our initial assumption about the frequency limit? Both answers are negative, as the following example will show.

Consider the sequence of 0's and 1's consisting of regularly increasing runs (see Chapter I, Eq. (4)).

$$0\ 1\ 00 \mid 11\ 00 \mid 0\ 111 \mid 0000 \mid 1111 \mid 0000 \mid 0\ 111 \mid 11 \cdots . \tag{29}$$

Here, the frequency of ones (and of zeros) has the limit $\frac{1}{2}$, as can easily be seen. The sequence (29) is, of course, not a collective since it obeys an explicitly given arithmetical rule; the condition of randomness is not fulfilled and we cannot expect that the argument that led to the Bernoulli theorem will be valid in this case. In fact, the statement is by no means true, although the frequency limit exists. Take, for instance, $n = 4$; that is, consider groups of 4 successive trials. At the beginning, the relative numbers $z = x/n$ (where x designates the number of one's in four observations) have various values: $\frac{1}{4}$, $\frac{2}{4}$, $\frac{3}{4}$, 0, 1, etc. But if we go farther in sequence (29), we come to runs of increasing length. Within a run of 1000 ones, we have 250, or at least 249, successive groups of length 4, to each of which corresponds the value $z = 1$. This behavior continues as we go farther and farther; there will be almost exclusively groups with $z = 0$ and groups with $z = 1$ (both with approximately the same frequency) and almost no groups with an intermediate value of z. Exactly the same is true if we choose any other (larger) length n instead of $n = 4$; we need only go farther ahead in the infinite sequence. The fact is that in the sequence (29), where the limiting frequency of ones is $\frac{1}{2}$, the chance of encountering a group, whatever its length n, with a z-value near to $\frac{1}{2}$ is nil; only groups with $z = 0$ or with $z = 1$ have a finite chance. According to the Bernoulli theorem, just the opposite should happen: the chance of z-values close to $\frac{1}{2}$ should be nearly 1 for groups with sufficiently large n.

A less artificial example of the same type can be observed by means of a table of square roots. If the table gives 7 decimals places, say, and we consider the sixth figure after the decimal point and decide to write a "0," if this figure equals 0, 1, 2, 3, or 4 and a "1" if it equals 5, 6, 7, 8, or 9, we obtain a sequence of 0's and 1's which behaves quite similarly to the sequence (29). See more in v. Mises [22], p. 111 ff.[1]

Thus, it is seen that the Bernoulli theorem states an important particular property of random sequences or collectives, a property which goes well beyond the mere existence of frequency limits, p, q. The property can roughly be described in this way: In a random sequence with $q = \frac{1}{2}$, the frequency limit $q = \frac{1}{2}$ of ones is brought about in such a way that almost all groups of sufficient length, n, include approximately one-half ones and one-half zeros. On the other hand, in the sequence (29), the frequency limit $q = \frac{1}{2}$ is such that whatever the length n, groups consisting exclusively of ones alternate with groups consisting of zeros only. Note that without the existence of such a Bernoullian property, which implies that a certain stabilization and compensation takes place in most groups of length n, no statistical regularities would have been observed in gambling or in other social or natural phenomena.[2] On the other hand, randomness is not a *necessary* condition for the validity of Bernoulli's theorem, as seen by the example of the regular sequence 010101 ... . This shows that Bernoulli's theorem cannot serve as a substitute for the postulate of randomness.[3]

We finish the discussion by showing (see also Appendix Two, p. 45) that the Bernoulli theorem reduces to a purely arithmetical property of numbers if Laplace's classical (or any "modernized classical") definition of probability is used. If 2^n different binary numbers, each consisting of n digits equal to either 0 or 1, are considered, then for increasing n an increasing percentage of these numbers will contain nearly as many zeros as ones. That is all that can be concluded.[4] If one uses a probability concept that does not relate to the frequency of an event, neither the

[1] This example illustrates how carelessly problems of probability theory have been treated by many mathematicians. In various textbooks of the calculus of probability, sequences of the above kind have been actually used as examples for the application of the theory, although they directly contradict the condition of randomness, considered one way or the other as essential by most workers in the field. They also contradict Bernoulli's theorem. However, the frequency limits for the 10 digits at the ninth place of a table of square roots exists (see a proof in v. Mises [21], p. 184). In the table of logarithms, used by H. Poincaré in a discussion of this kind, not even the limits of the relative frequencies exist.

[2] Compare, in this connection, part (i) of Appendix Three which concerns "rapid convergence."

[3] We have seen in Chapter I, Section 8, that within a collective the multiplication theorem holds; this condition, in addition to the existence of a chance, is sufficient for the validity of Bernoulli's theorem. In the regular sequence above, the multiplication theorem, of course, does not hold.

[4] This, for $p = \frac{1}{2}$, is the Borel Theorem of page 42. A similar, purely arithmetical statement holds for any p (see bottom of p. 42).

Bernoulli theorem, nor any of its generalizations, lead to any statement concerning the frequencies in a sequence of trials.

Problem 12. How often must a correct die be thrown to ensure a 95% probability (according to Tchebycheff's inequality) that the number of sixes does not deviate by more than 5% from its expected value?

Problem 13. Make an arithmetical statement that corresponds to the Bernoulli theorem if $p = \frac{1}{3}$.

5. The Approximation to the Binomial Distribution in the Case of Rare Events. Poisson Distribution

5.1. *Problem and solution.* We return to the Bernoulli distribution $p_n(x)$ given by (11) and observe that for large values of n it cannot easily be evaluated since it contains $n!$ and $(n - x)!$; further, the computation of probabilities of individual results does not give satisfactory insight into the behavior of the distribution as a whole. In his *Théorie Analytique des Probabilités* Laplace studied the asymptotic behavior of $p_n(x)$ in the case in which the mean value nq and the variance npq become large together with n. Laplace's result, which as we shall see, leads to the normal distribution as the approximation function, will be given in Chapter VI. Here, we consider another asymptotic result, mathematically much simpler, which will be of particular interest in connection with problems to be dealt with in the present chapter.

There are, in fact, cases in which the Laplace approximation is not satisfactory at all. Let us consider the following example. The collective may consist of successive childbirths, and the "event," occurring with probability q, may be the birth of triplets. Here q is, in order of magnitude, 10^{-4}. One may ask for the probability that 0 or 1 or 2 triplets are born in a succession of $n = 100{,}000$ births. The answer is supplied by Newton's formula (11) which gives, if we use $q = 0.0001$ and $n = 100{,}000$,

$$p_n(0) = 0.000045378, \qquad p_n(1) = 0.00045382, \qquad p_n(2) = 0.0022693.$$

If we compute the corresponding values from the Laplace approximation, (35) of Chapter VI, we find

$$p_n(0) = 0.000850, \qquad p_n(1) = 0.00220, \qquad p_n(2) = 0.00514.$$

These figures are certainly not a satisfactory approximation to the correct values. The reason is that, as noted above, a basic assumption in deriving Laplace's formula is that the variance npq becomes infinite with n. In the present example, with $n = 10^5$, $q = 10^{-4}$; the variance $s_n{}^2 = npq$ has only about the value 10 and this is not a "large" quantity. The extremely small q restricts the magnitude $s_n{}^2$ in spite of the large value of n. A small q means that the event under consideration is a "rare event."

S. D. Poisson showed in his *Recherches sur la Probabilité des Jugements* how, in a simple way, one can find a very good approximation for $p_n(x)$ in the case of rare events. *Poisson's law*, which will be derived presently, is of great theoretical interest and has a wide range of application in social statistics as well as in problems of theoretical physics (see Section 5.2).

Let us introduce, in the Bernoulli formula, the mean value $a = nq$ and write (11) in the form

$$p_n(x) = \frac{n(n-1)\cdots(n-x+1)}{x!}\left(1-\frac{a}{n}\right)^{n-x}\left(\frac{a}{n}\right)^x$$

$$= \frac{a^x}{x!}\left(1-\frac{a}{n}\right)^n \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)}{\left(1-\frac{a}{n}\right)^x}. \tag{30}$$

If n increases while a and x are kept constant, the last fraction consists of a finite number x of factors, each of which tends toward unity as $n \to \infty$. The product, therefore, goes to 1 too. On the other hand, it is well known that

$$\lim_{n\to\infty}\left[\left(1-\frac{a}{n}\right)^n\right] = e^{-a}.$$

Thus

$$\lim_{\substack{n\to\infty \\ x,\, a \text{ fixed}}} p_n(x) = \frac{a^x e^{-a}}{x!} = \psi(x). \tag{31}$$

The right-hand expression can now serve as an approximation for $p_n(x)$ if n is large and x as well as $a = nq$ remain moderate in value. (The latter assumption implies that q is small, i.e., that we are dealing with a rare event.) In the above example, with $n = 100{,}000$ and $q = 0.0001$, that is $a = 10$, the Poisson expression (31) gives the following results for $x = 0, 1, 2$:

$$\psi(0) = 0.00004540, \qquad \psi(1) = 0.0004540, \qquad \psi(2) = 0.002270.$$

The approximation to the correct value, given on p. 177 is seen to be excellent.[1]

5.2. *Discussion and applications.* The goodness of the Poisson formula (see Table V)

$$p_n(x) \sim \frac{a^x e^{-a}}{x!} \tag{31'}$$

as an approximation to (11), can be estimated if we return to Eq. (30) and replace the product

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)$$

by its lower and upper bounds $\left(1-\frac{x-1}{n}\right)^{x-1}$ and 1, thus obtaining (for $x > 1$) the inequalities

$$\frac{a^x}{x!}\left(1-\frac{a}{n}\right)^{n-x}\left(1-\frac{x-1}{n}\right)^{x-1} < p_n(x) < \frac{a^x}{x!}\left(1-\frac{a}{n}\right)^{n-x}. \tag{32}$$

The factor $(1 - a/n)^n$ occurring here can be estimated in the following way. We start from well-known inequalities, used, for example, in proving the existence of the limit of $(1 - 1/n)^n$, which equals e^{-1}. These inequalities are

$$\left(1-\frac{1}{z}\right)^{z} < e^{-1} < \left(1-\frac{1}{z}\right)^{z-1} \quad \text{for any} \quad z > 1.^2 \tag{33}$$

Replacing z by n/a, we obtain from (33)

$$\left(1-\frac{a}{n}\right)^{n/a} < e^{-1} < \left(1-\frac{a}{n}\right)^{(n-a)/a}. \tag{34}$$

This is equivalent to

$$\left(1-\frac{a}{n}\right)^{n} < e^{-a} \quad \text{and} \quad \left(1-\frac{a}{n}\right)^{n} > e^{-a}\left(1-\frac{a}{n}\right)^{a}. \tag{34'}$$

[1] To see that $\psi(x) = (a^x/x!)e^{-a}$ is not the probability of a limit set, put $q = a/n$, $n = [a] + 1, [a] + 2, \ldots$, and denote by A_n the set of all sequences whose first n terms show the event x times (a sum of basic sets). Consider $A_n, A_{n+1}, A_{n+2}, \ldots$; this sequence converges and its limit is the set A of all sequences with exactly x events. The probability of A is, however, equal to zero and not to $\psi(x)$; $\psi(x)$ is an analytic approximation to $p_n(x)$ valid under the conditions of the text.

[2] Here $(1 - 1/z)^z$ is monotonically increasing, and $(1 - 1/z)^{z-1}$ is monotonically decreasing and both have the limit e^{-1} as $z \to \infty$.

Consequently, Eq. (32) yields

$$\frac{a^x}{x!}\left(1-\frac{a}{n}\right)^a e^{-a}\left(1-\frac{a}{n}\right)^{-x}\left(1-\frac{x-1}{n}\right)^{x-1} < p_n(x) < \frac{a^x}{x!}e^{-a}\left(1-\frac{a}{n}\right)^{-x} \tag{35}$$

or, using the notation $\psi(x)$ introduced in (31):

$$\left(1-\frac{a}{n}\right)^{a-x}\left(1-\frac{x-1}{n}\right)^{x-1} < \frac{p_n(x)}{\psi(x)} < \left(1-\frac{a}{n}\right)^{-x}. \tag{35'}$$

It is seen that both the first and the last expressions tend to 1 as $n \to \infty$ and the approximation is good if a/n is small and x not too large. We note for later use that the asymptotic equality of $p_n(x)$ and $\psi(x)$ as determined by (35') holds even if a tends toward infinity, as long as a^2/n tends toward zero (as seen from $(1-a/n)^a = [(1-a/n)^{n/a}]^{a^2/n} \to (e^{-1})^{a^2/n} \to 1$). The above inequalities determine an upper bound for the error committed in replacing $p_n(x)$ by $\psi(x)$. Applied to the example discussed before ($n = 10^5$, $a = 10$) Eq. (35') gives for $x = 2$, to the left $0.9999^8 \times 0.99999 = 0.9992$ and to the right $0.9999^{-2} = 1.0002$ as lower and upper bounds of the ratio $p_n(2)/\psi(2)$. The actual value of this ratio in our example is 0.9997.

The so-called *Brownian movement*, which we shall discuss shortly, furnishes an example in which Poisson's formula is applied to a problem of molecular physics. If a liquid or gas under appropriate conditions is observed through a high-powered microscope, one sees small suspended particles (dust or smoke particles, colloidal or crystalline particles, that have been added intentionally) move around in a haphazard way. In order to check the random nature of this phenomenon, one may count the number of particles that are present at a certain instant in a well-defined portion A of the space occupied by the fluid (the field of vision of the microscope) and repeat the count many times in succession. In this way, one may determine the frequencies of the event that the region A is occupied by 0, 1, 2, ..., x particles (see Problem 17).

The number n of particles present in the fluid is very large. Since the space A is small compared with the total space occupied by the fluid, the individual probability q for a particle to be found in A is very small, and the expected number nq of particles observed in A at any one time remains moderate. Thus, the conditions under which Poisson's formula approximates the Bernoulli formula are fulfilled.

In order to apply the theory, one has first to estimate the value of $nq = a$, the expected number of particles present in the space of observation. If a sufficiently large number m of observations have been made and altogether M particles have been counted, one may assume that M/m represents a rough estimate of the value of a.[3] The approximate probability of finding x particles is then $e^{-a}a^x/x! = \psi(x)$, and $m\psi(x)$ is the expected number of those cases in which x particles were counted. These expectations are the quantities to be compared with the observed frequencies. For instance, if in $m = 500$ observations, 1500 particles in toto have been counted, then a is approximately equal to 3 and $500e^{-3}3^2/2! = 112$ should approximate the number of observations with $x = 2$. The object of such an investigation is, of course, to decide whether the simultaneous appearance of Brownian particles in a given space of observation can be considered as a Bernoulli problem, or, more generally, whether the movement of the Brownian particles has the character of a random phenomenon.

Apart from its meaning as an asymptotic expression for (11), the function $e^{-a}a^x/x!$ can be considered as the definition of a probability distribution for the discontinuous chance variable x in the range 0 to ∞, since

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

The distribution $\psi(x)$ (which depends only on one parameter a) is known as the *Poisson distribution*. Some basic properties of $\psi(x)$ have been given in Chapter III, Section 6. A few applications follow in the problems of this section. Further applications will be given in some of the following sections of the present chapter. More general considerations will be found at the end of Chapter VI and in Appendix Four.

Problem 14. The probability of a multiple birth of more than 3 children is 1.5×10^{-6}. What is the probability that in 10^6 cases 1, 2, 3 such multiple births occur?

Problem 15. Give upper and lower bounds for the probability $p_n(3)$ in the case of problem 14.

Problem 16. In a certain community an average of one suicide per week occurs. What is the expected number of weeks with 3 or more suicides, within one year?

Problem 17. The following number of Brownian particles have been observed in 1583 observations:

0 particle .. 381 times,	4 particles .. 67 times,
1 particle .. 568 times,	5 particles .. 28 times,
2 particles .. 357 times,	6 particles .. 5 times,
3 particles .. 175 times,	7 particles .. 2 times.

[3] See however Problem 17 where a can be estimated more accurately from the data.

Estimate from the data the parameter of the corresponding Poisson distribution. Compute from the Poisson formula the expected numbers of observations with $x = 0, 1, 2, ..., 7$ and compare them with the actual counts.

6. The Negative Binomial Distribution

We introduce briefly another useful discontinuous distribution related to the binomial distribution. Consider n repeated trials with probabilities p and q for 0 and 1 (failure and success). We ask for the probability that in our n trials the first success occurs at trial number $(k + 1)$, $[k = 0, 1, ..., (n - 1)]$. This will happen if there are first k "failures" and then a "success"; the probability is $p^k q$, $k = 0, 1, ..., (n - 1)$; finally, it can happen, with probability p^n, that all n trials result in failure. Thus, we obtain the distribution

$$\begin{aligned} g_n(x) &= qp^x, \qquad x = 0, 1, ..., n - 1, \\ &= p^n, \qquad x = n. \end{aligned} \tag{36}$$

We verify that $\Sigma_{x=0}^{n} g_n(x) = q\ \Sigma_{x=0}^{n-1} p^x + p^n = q(1 - p^n)/(1 - p) + p^n = 1$.

The limit of this distribution as $n \to \infty$ is the *geometric distribution*

$$g(x) = qp^x, \qquad x = 0, 1, ..., \tag{37}$$

the probability that the first success occurs at the trial numbered $x + 1$, where $x = 0, 1, ...$. The reader will easily verify that $\Sigma_{x=0}^{\infty} g(x) = 1$, $\Sigma_{x=0}^{\infty} xg(x) = p/q$.

The probability distribution (37) can be considered a special case of the so-called *negative binomial distribution*. We ask for the probability that, within n trials, the rth "success" is at the trial numbered $k + r$, $k = 0, 1, ..., n - r$. [Here, $r = 1$ corresponds to (36)]. This event occurs if, among the first $k + r - 1$ trials there are $(r - 1)$ successes and k failures, and then if the next trial results in a success. The first probability is $\binom{r+k-1}{k} p^k q^{r-1}$ and the second is q; thus, the required probability is

$$\binom{r + k - 1}{r - 1} p^k q^r, \qquad k = 0, 1, ..., n - r.$$

The sum of these $n - r + 1$ probabilities is not unity since it may also happen that the n trials result in fewer than r successes, namely, in $r - y$ successes, where $y = 1, 2, ..., r$. The respective probabilities are $\binom{n}{r-y} q^{r-y} p^{n-r+y}$, and if these r probabilities are added to the former ones, the sum is unity. We are, however, again interested in the case where

$n \rightarrow \infty$, while r remains finite. Then, it can be seen easily that the sum of these last probabilities tends to zero and we obtain the distribution

$$b_r(x) = \binom{r+x-1}{r-1} p^x q^r, \qquad x = 0, 1, 2, \ldots. \tag{38}$$

Here in the symbol $\binom{r+x-1}{r-1}$, the upper line depends on the variable x.

We will rewrite this binomial symbol by generalizing its original definition. For m and s positive integers, $s \leqslant m$, we have $\binom{m}{s} = \frac{m(m-1)\cdots(m-s+1)}{1 \cdot 2 \cdots s} = \frac{m^{(s)}}{s!}$, where the factorial symbol $m^{(s)}$ stands for the product of the s factors $m\ (m-1) \cdots (m-s+1)$. Now, let a be a positive integer and define

$$\begin{aligned}\binom{-a}{s} &= \frac{-a(-a-1)(-a-2)\cdots(-a-s+1)}{s!}\\ &= (-1)^s \frac{a(a+1)(a+2)\cdots(a+s-1)}{s!}\\ &= (-1)^s \frac{(a+s-1)^{(s)}}{s!} = (-1)^s \binom{a+s-1}{s}.\end{aligned} \tag{39}$$

We then obtain $\binom{r+x-1}{r-1} = \binom{r+x-1}{x} = (-1)^x\binom{-r}{x}$, hence

$$b_r(x) = (-1)^x \binom{-r}{x} p^x q^r = \binom{-r}{x} (-p)^x q^r, \qquad x = 0, 1, \ldots. \tag{40}$$

We prove that the sum of the $b_r(x)$ equals one, by using the binomial theorem for $(1-p)^{-r}$. Infact,

$$(1-p)^{-r} = \sum_{x=0}^{\infty} \binom{-r}{x} (-p)^x. \tag{40'}$$

Multiplying both sides of this equation by q^r, we obtain

$$(1-p)^{-r} q^r = 1 = \sum_{x=0}^{\infty} \binom{-r}{x} (-p)^x q^r = \sum_{x=0}^{\infty} b_r(x). \tag{40''}$$

In formula (40′), r is not necessarily an integer. Originally, the r in the problem of the rth success arising at trial numbered $k+r$ was a positive integer, but (40″) remains meaningful for any non-negative r, hence in Eq. (38) or (40), *r need only be* $\gtrsim 0$. The distribution (40) or (38) is called the negative binomial distribution. If r is a positive integer, $b_r(x)$ is the probability that the rth success arises at trial $x + r$, where $x = 0, 1, \ldots$.

In Chapter V, Eq. (57), we shall compute the mean value and variance of $b_r(x)$.

Problem 18. In the negative binomial distribution let $p \to 0$, $r \to \infty$, in such a way that $rp = a$ remains fixed. Prove that $b_r(x)$ converges toward the Poisson distribution $(a^x/x!)\, e^{-a}$.

C. Some Problems of Non-Independent Events (Sections 7-9)

7. A Problem of Runs

7.1. *Problem and solution for small n.* Let us take as a starting point for the discussion of the theory of runs the almost superstitious attitude of many people toward the occurrence of a run, that is, of a sequence of equal label values in a random series of observations. People believe that there really should not be any long runs in a random series of two alternative label values with the probabilities $\frac{1}{2}$. If in a game of heads and tails, heads comes up five times in succession, or if in a maternity ward, five girls are born in succession, there is always somebody who ventures to say: "This is against the laws of probability." In terms of gambling, this would mean that it is particularly advantageous to bet on "tails" if the preceding turns make up a run of "heads" of some length. In our terminology this amounts to assuming that coin tossing is, after all, not a collective and that a clever place selection may change the probabilities.

The German philosopher, K. Marbe, tried to develop a system based on the idea that long runs contradict probability calculus. He investigated painstakingly the birth records of four cities, each record containing about 50,000 entries, and searched for sequences of male or female newborn children. The longest run he found consisted of 17 entries of the same sex in a row. He came to the conclusion that there is something in the popular belief that after 17 girls have been born in succession the next child must be a boy.

As already indicated, Marbe's conclusion questions the randomness of a particular observed sequence (birth records). The approach of probability calculus in this case is to draw conclusions following from assumed randomness and to compare them with the facts. We therefore start with the assumption that before us we have a collective K, with two label values, m and f or 0 and 1, which possess the respective probabilities

p and q, and try to find the probability of a run in this infinite random sequence.

This "probability of a run" is yet to be defined; that is, a collective must be constructed whose label values are connected in a unique way with the runs occurring in the original alternative K. Let us first state the meaning we assign to the term run: a run of length m is a sequence of m equal label values (say, zeros) immediately preceded and followed by one opposite label value (one). The following finite sequence

$$1\,1\,0\,0\,0\,1\,1\,0\,0\,1\,1\,1\,1\,0\,1\,1\,1\,1\,0\,0\,0 \qquad (41)$$

contains one zero-run of length 1, one of length 2, one of length 3, one one-run of length 2, and two of length 4. Note that the double-one at the start and the triple-zero at the end are not counted as runs in this definition.

As often before (see, e.g., Section 3.1) we consider now a collective K', the elements of which are successive groups of n elements of the original collective; as in the problem of repeated trials, K' is constructed by combining the results of n place selections operated on K. The n-dimensional label of an element of K' consists of a sequence of n zeros and ones, and if it consists of α zeros and β ones, its probability is $p^\alpha q^\beta$ with $\alpha + \beta =$ n. Now, however, we are not merely interested in the number of "ones" in a label of K' but in the succession of the 0's and 1's, in particular, with respect to runs. Each element of K', being a sequence of zeros and ones, will include a certain number of runs of various lengths $m = 1, 2, ..., n - 2$. (We consider simultaneously runs of zeros and ones.) In the infinite sequence K', those elements of K' which include exactly x runs of length m will have a certain limiting frequency. Since this limit depends on n and m, we call it $P_n^{(m)}(x)$, that is, *the probability of obtaining x runs of length m within a sequence of n trials.*[1]

To find $P_n^{(m)}(x)$, we have to carry out a specific mixing operation on K'; that is, we have to sum up all those $p^\alpha q^\beta$-values that belong to labels of K' possessing exactly x runs of length m.

Take as an example $n = 5$. The collective K' admits, altogether, $2^5 = 32$ different label values (all possible binary fractions with 5 places), half of which are listed here together with their respective probabilities:

(1)	00000	p^5	(9)	01000	p^4q
(2)	00001	p^4q	(10)	01001	p^3q^2
(3)	00010	p^4q	(11)	01010	p^3q^2

[1] It would be more in keeping with our usage to write $p_n^{(m)}(x)$; we use, however, capital P, in order to distinguish the present probability from $p_n(x)$ of Eq. (11).

(4)	00011	p^3q^2	(12)	01011	p^2q^3
(5)	00100	p^4q	(13)	01100	p^3q^2
(6)	00101	p^3q^2	(14)	01101	p^2q^3
(7)	00110	p^3q^2	(15)	01110	p^2q^3
(8)	00111	p^2q^3	(16)	01111	pq^4

The remaining 16 cases are obtained by interchanging 0 and 1, and p and q. Suppose we want to find $P_5^{(2)}(1)$. The labels in the table containing exactly one run of length 2 are #7, 10, 13, 14, with the probabilities p^3q^2, p^3q^2, p^3q^2, p^2q^3, respectively. If we add the corresponding values due to interchanging 0 and 1, we have

$$\begin{aligned} P_5^{(2)}(1) &= (3p^3q^2 + p^2q^3) + (3p^2q^3 + p^3q^2) \\ &= 4p^2q^2(p + q) = 4p^2q^2. \end{aligned} \tag{42}$$

In the same way, e.g., $P_5^{(3)}(1)$ is found as p^2q^3 and similarly all other probabilities may be obtained. Should we use a definition of a run that includes the cases of m equal figures at the beginning or at the end, the values of $P_n^{(m)}(x)$ would change slightly.

7.2. *Expected number of runs.* This method of computation by tabulation of the possible labels is only applicable if n is small. For large n, we cannot give a practical way to compute $P_n^{(m)}(x)$. However, a quantity which is of greater interest than the probability of each single x is the mean value or expected value of x, which is defined as

$$\begin{aligned} E_n^{(m)}[x] &= 1 \cdot P_n^{(m)}(1) + 2 \cdot P_n^{(m)}(2) + 3 \cdot P_n^{(m)}(3) + \cdots \\ &= \sum_{x=0}^{\infty} xP_n^{(m)}(x). \end{aligned} \tag{43}$$

We affix the n, m to the symbol E to indicate that E depends on the two parameters n, m. Note, also, that the sum on the right-hand side of (43) consists of a finite number of terms since $P_n^{(m)}(x) = 0$ for $x > (n - 2)/m$. The following argument will lead to the value of $E_n^{(m)}[x]$.

Imagine a run of length m occurring in a sequence of length n. This implies filling m consecutive places with zeros, say, and framing them with two ones (see Fig. 10). The probability that an n-dimensional label includes such a run must contain the factor q^2p^m. If we sum the probabilities of all possible labels which include such a run of m zeros in a prescribed position (that is, one starting at $k + 1$), we may factor out the expression q^2p^m; this expression is multiplied by a sum of products, each containing $n - (m + 2)$ factors p, q and each possible arrangement of p's and q's occurs just once. Each such product is one of the 2^{n-m-2} terms of which the $(n - m - 2)$th power of $(p + q)$ is composed. Therefore, the factor that multiplies q^2p^m equals $(p + q)^{n-m-2}$, and since $p + q = 1$, it equals one. In other words, q^2p^m is the probability of all possible sequences of length n which exhibit a run of zeros of length m at a preassigned "place," regardless of the magnitude of $n \geqslant m + 2$. In the same way, p^2q^m is the probability of a run of m ones at a given place.[2] (By "place" we mean here the $m + 2$ preassigned spots from k to $k + m + 1$.)

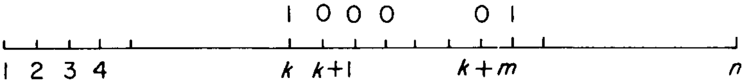

FIG. 10. Run of length m.

In the above consideration, the 1 preceding the run of zeros (or the 0 preceding the run of ones) occupies the kth place in our sequence, where k may be any one of the numbers 1, 2, ..., $n - m - 1$. Let us therefore consider the expression

$$(n - m - 1)(p^2q^m + q^2p^m). \tag{44}$$

This is the sum of the probabilities of the following events:

the sequence has a run of m zeros or a run of m ones beginning at $k = 1$

the sequence has a run of m zeros or a run of m ones beginning at $k = 2$

. . . . (45)

. . . .

the sequence has a run of m zeros or a run of m ones beginning at $k = n - m - 1$.

These events are not mutually exclusive, since it is possible that the same sequence constitutes an event that has been listed more

[2] This is true not only for a run; generally, $p^{\alpha}q^{\beta}$ is the probability of having α zeros and β ones at preassigned places with any labels whatever on the remaining $n - (\alpha + \beta)$ places; $p^{\alpha}q^{\beta}$ is a marginal probability.

than once. Let $n = 14$, $m = 4$, and consider the arrangement $\underline{|0\ 1\ 1\ 1\ 1\ 0|}\ 0\ \overline{|0\ 1\ 1\ 1\ 1\ 0|}\ 1$ of probability p^5q^9; this probability will be counted twice, once in the computation of the marginal probability p^2q^4 which corresponds to $k = 1$, the second time in the computation of that p^2q^4 which corresponds to the run starting at $k = 8$. In general, it is seen that those sequences which include one and only one run (always of length m) appear only once in the enumeration (45), those with two runs will appear exactly twice, and those with x runs are listed x times. The sum of the probabilities of all events (45) is therefore equal to: the probability of one run plus twice the probability of two runs plus ... . In other words, we have proved that the expression (44) equals the right-hand side of Eq. (43). Thus, the mean value a_m or the expectation of x is

$$a_m = \sum_{x=0}^{\infty} xP_n^{(m)}(x) = E_n^{(m)}[x] = (n - m - 1)(p^2q^m + q^2p^m). \tag{46}$$

If n is a large number, the length m of runs which are of practical interest will be small compared with n. In this case, a_m will be approximately equal to $n(p^2q^m + q^2p^m)$. If $p = q = \frac{1}{2}$, this yields

$$a_m \sim n(\tfrac{1}{2})^{m+1} \tag{47}$$

as the expected number of runs of length m in a sequence of n observations with two possible labels of equal probability.

Let us now compare the results of Marbe's observations with the expectations computed under the assumption that the sequence of male or female births is a random series. The p, q in this sequence are taken from the observed frequencies as

$$\begin{aligned} p &= 0.51099 \qquad \text{for a male birth,} \\ q &= 0.48901 \qquad \text{for a female birth.} \end{aligned} \tag{48}$$

Marbe analyzed 4 sequences, n being in each case 49,152. In the following tabulation, the number of runs observed are given together with their arithmetical means (a.m.) for $m = 1, 5, 10, 17, 18$. The value of the corresponding a_m, computed according to (46), is in the third column.

One sees that there is excellent agreement between the observed averages and the expected numbers of runs derived under the assumption of randomness in the sequence of male and female births. Still one might feel that there should be some more explanation of the fact that no run of more than 17 elements was observed, since theoretically runs of any given length m are possible and must occur in a sequence of sufficient length n. If the distribution $P_n^{(m)}(x)$ were known, we could compute $P_n^{(18)}(0)$ and, in this way, obtain information about how probable the complete lack of runs of length 18 is for the given value of n.

Length of run	Number of runs observed	Computed value of a_m
$m = 1$	12,305	
	12,028	
	12,154	
	12,136	
a.m. =	12,156	12,282
$m = 5$	780	
	735	
	813	
	761	
a.m. =	772	768
$m = 10$	23	
	21	
	31	
	22	
a.m. =	24.3	24.3
$m = 17$	0	
	0	
	1	
	0	
a.m. =	0.25	0.20
$m = 18$	0	
	0	
	0	
	0	
a.m. =	0	0.10

7.3. *Estimates for large n.* Instead of computing $P_n^{(18)}(0)$, we shall determine a lower bound for this quantity by means of Tchebycheff's inequality. We need the variance $s_m{}^2$ of $P_n^{(m)}(x)$.

Consider again a sequence of length n, but now assume right away that m is negligibly small compared with n. This amounts to assuming that on a line of length n the space occupied by a run is so small that one may speak of the run as being located at one of the n integral points of the line. The probability that a sequence of length n includes 2 separate zero-runs of length m located at certain places can be computed analogously to the computation given above for the probability of a sequence with one run. Since now twice $(m + 2)$ given places are occupied by specified figures ($2m$ by zeros and 4 by ones), the probability will be $p^{2m}q^4 = (p^m q^2)^2$. In the same way, $(p^2q^m)^2$ is the probability for 2 one-runs of length m at given places, and the product $(p^2q^m)(q^2p^m)$ represents the probability of having a one-run at one place and a zero-run at another place. Hence, if two places for two runs are specified, the probabilities for all sequences that include a run of either kind at each of those places will be $(p^2q^m + q^2p^m)^2$.

Now, if the length m of a run is negligible compared to n, there are $n(n - 1)/2$ possibilities of placing two runs among the n places in the sequence. Therefore, we consider the expression analogous to (44)

$$\frac{n(n-1)}{2}(p^2q^m + q^2p^m)^2, \tag{49}$$

which is the sum of the probabilities of all arrangements which contain at least two runs of length m.

These events are not mutually exclusive: a sequence that contains exactly x runs of length m is counted in (49) as often as it is possible to select a group of two runs out of x runs, that is, $x(x - 1)/2$ times. Hence, we may write (for large n)

$$\frac{1}{2}\sum_{x=0}^{\infty} x(x-1)P_n^{(m)}(x) = \frac{n(n-1)}{2}(p^2q^m + q^2p^m)^2. \tag{50}$$

Now,

$$\sum_{x=0}^{\infty} x(x-1)P_n^{(m)}(x) = \sum_{x=0}^{\infty} x^2P_n^{(m)}(x) - \sum_{x=0}^{\infty} xP_n^{(m)}(x). \tag{51}$$

On the right-hand side of this equation, the first term equals $s_m{}^2 + a_m{}^2$. The second term is a_m. Upon introducing these expressions on the left-hand side of (50) and replacing $\frac{1}{2}n(n - 1)(p^2q^m + q^2p^m)^2$ by $\frac{1}{2}a_m{}^2$ [by Eq. (46) using the fact that n is large compared to m] we arrive at

$$\tfrac{1}{2}(s_m{}^2 + a_m{}^2 - a_m) = \tfrac{1}{2}a_m{}^2 \qquad \text{or} \qquad s_m{}^2 = a_m = n(p^2q^m + q^2p^m). \tag{52}$$

Let us now set up Tchebycheff's inequality (Chapter III, Section 1.2) for an interval chosen in such a way that we obtain a lower bound for

$P_n^{(18)}(0)$. For $m = 18$, we have $a_m = 0.1$ and $s_m{}^2 = 0.1$. If, in Tchebycheff's inequality, we choose $X = 0.9$ for the half-length of the interval, the condition for $x - a_m$ to fall into that interval is simply that x must be zero, since

$$| x - 0.1 | < 0.9 \tag{53}$$

is to be satisfied by an integer x. Tchebycheff's inequality leads therefore to

$$P_n^{(18)}(0) \geqslant 1 - \frac{s_m{}^2}{X^2} = 1 - \frac{0.1}{0.81} = 0.88 \tag{54}$$

If we take $X = 1.9$, the right-hand side of (54) becomes 0.97.

Thus, the result $x = 0$ (no run of length 18) had to be expected with a probability of at least 88%. The probability of having more than one run of length 18 is smaller than 3%. It cannot be said that the observations lead to any conclusion "in disagreement with probability calculus."

The same method which has been used here for finding the variance can be used to find all higher moments of $P_n^{(m)}(x)$. It is then possible (see Section 9.4 of this chapter) to determine for large n the asymptotic value of $P_n^{(m)}(x)$. The result is that asymptotically for large n, $P_n^{(m)}(x)$ approaches the distribution $(a^x/x!)e^{-a} = \psi(x)$, where $a = \lim_{n\to\infty} n(p^2q^m + q^2p^m)$ remains finite as $n \to \infty$. In other words, the probability that in n trials there appear x runs of length m approaches $\psi(x) = (a^x/x!)e^{-a}$ as $n \to \infty$, where a is the expected value of x. For example, if we return to the birth records, for $m = 16$, $a = 0.391$, we obtain $\psi(0) = 0.68$, $\psi(1) = 0.26$, $\psi(2) = 0.05$. In the four sequences of roughly 50,000 observations, no run of length 16 was observed, that is in each of 4 observations, an event occurred which had a probability of 0.68. For $m = 18$, $a = 0.098$, we obtain $\psi(0) = 0.907$, hence a 91% probability for no run of this length,[3] and $\psi(1) = 0.089$, $\psi(2) = 0.004$.

Problem 19. What is the largest run that can be "fairly" expected (having an expectation approximately equal to 1) in a sequence of 500 tossings of an unbiased coin? How many runs of length 6 can be expected in this sequence? Find a lower bound for the probability that one to six runs of length 6 occur.

Problem 20. What expression should be added on the right-hand side of Eq. (46) if a succession of m equal label values at the end and at the beginning of the sequence is also counted as a run of length m? Show that this addition is unimportant if n is large.

[3] R. v. Mises, "Das Problem der Iterationen." *Z. Angew. Math. Mech.* 1 (1921), pp. 298–307.

Problem 21. By tabulating the possible results, in the case $n = 5$, compute all probabilities for runs of length $m = 2$ and from these probabilities the expectation of the number x of runs. Compare this result with the value found from Eq. (46) and with the one adjusted according to the preceding problem.

Problem 22. How many runs of m ones, or twos, ... or sixes can be expected in n tossings of a die? At what length of the game with an unbiased die can a run of length 6 be "fairly" expected?

Problem 23. Give the exact expression (finite m) for the variance of x in the case of an alternative with probabilities p, q and compare the result with the outcome of the computation for $n = 5$ (Problem 21).

8. Arbitrarily Linked Events. Basic Relations

In this section we shall derive the basic relations which hold for n *compatible*, or *arbitrarily linked events*.[1] We shall see in Sections 9 and 10 that many interesting problems, among them the problem of runs studied in the preceding section, can be considered from this viewpoint.

Consider a collective of a label space S; let $A_1, A_2, ..., A_n$ be certain subsets of S, and $p_1, p_2, ..., p_n$ the respective probabilities, derived from the given probability distribution over S. The A_i need not be particularly simple events. Think, for example, of k drawings (with replacement) out of an urn which contains m different kinds of balls numbered 1 to m; S consists of m^k points with the k coordinates of each point each being one of the numbers 1, 2, ..., m. Then, p_1, for example, may be the probability of the result "1" on the first trial and "6" on the third trial, that is A_1 contains all points of S whose first coordinate equals "1" and whose third coordinate equals "6"; next, A_2 might be the event with the result "1" on first, third, and fifth trials, and so on. The A_i may overlap in an arbitrary way. The label space need not be discrete; S might be a continuous space, and $A_1, A_2, ..., A_n$ subsets of S. Note that this is a theory of arbitrarily linked *events* only, i.e., the ith event is the set A_i, and the alternative is $A_i' = S - A_i$. Each point of S is characterized by n signs:$\epsilon_1, \epsilon_2, \epsilon_3, ..., \epsilon_n$ where ϵ_i means either A_i or A_i', either 1 or 0.

We denote by A_{ij} the intersection of the sets A_i and A_j and by p_{ij} the

[1] H. Geiringer, "Sur les variables aléatoires arbitrairement liées." *Rév. Math. Union Interbalcanique* **2** (1938), pp. 1–26; "Probability theory of arbitrary events," *Ann. of Math. Statist.* **9** (1942), pp. 260–271, M. Fréchet, *Les probabilités associées à un systéme d'événements compatibles et dépendants.* Actualités scientifiques et industrielles, Paris, 1940, 1942, and 1943. Contributions to the more elementary theory are also due to Gumbel, and to the deeper problems to Loève.

probability of A_{ij}, that is, of the simultaneous occurrence of A_i and A_j. (In our example $p_{12} = 0$ since the result of the third trial cannot be 1 as well as 6.) If $p_{ij} = 0$ the respective events are ***mutually exclusive.*** Similarly, p_{ijk} is the probability of the joint occurrence ofA_i, A_j, A_k and finally, $p_{12...n}$ that of the joint occurrence of all n events. There are n probabilities p_i, $\binom{n}{2}$ probabilities p_{ij}, etc. The scheme of mutually independent events in n successive trials, where for all subscripts:

$$p_{ij} = p_i p_j , \quad p_{ijk} = p_i p_j p_k , ..., \quad p_{12...n} = p_1 p_2 \cdots p_n \tag{55}$$

is also contained in our setup. [However, we do not say that, conversely, the numerical equalities (55) always imply independence (see discussion in Chapter I, Section 10). The A_i which we study are arbitrarily linked events which, in general, are not independent events as defined Chapter I, Section 9.2.]

We may represent the sample space S and the subspaces A_1, A_2, ..., A_n in a plane (see Fig. 11), and we assume, for example,

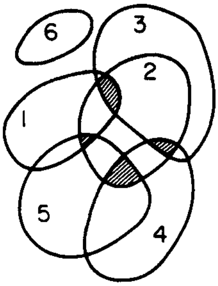

FIG. 11. Arbitrarily linked events.

that the probabilities of the various regions in the figure are proportional to the respective areas. Here $n = 6$; among the 15 probabilities p_{ij}, eight are different from zero, viz., p_{12}, p_{13}, p_{15}, p_{23}, p_{24}, p_{25}, p_{34}, p_{45} and the other seven are zero; the four probabilities p_{123}, p_{125}, p_{234}, p_{245} (corresponding to shaded domains) are different from zero, and the sixteen remaining ones are zero: all p_{ijkl}, etc., vanish.

Let $P_n(x)$ be *the probability that of the n events under consideration x take place.* Thus $P_n(x)$ corresponds to those points of S which are covered x times. Clearly,

$$P_n(x) = 0, \qquad x < 0 \quad \text{and} \quad x > n \tag{56}$$

and

$$\sum_{x=0}^{n} P_n(x) = 1. \tag{57}$$

In the case of Fig. 11, for example, with $n = 6$, a region characterized by $A_1 A_2 \bar{A}_3 \bar{A}_4 A_5 \bar{A}_6$ is "triply covered." It belongs to that part of S where three events take place and its probability is part of $P_6(3)$. We have obviously, in the case of the figure

$$P_6(1) = (p_1 + p_2 + \cdots + p_6) - 2(p_{12} + p_{13} + \cdots + p_{45}) + 3(p_{123} + \cdots + p_{245})$$

$$\begin{aligned} P_6(2) = {} & (p_{12} - p_{123} - p_{125}) + (p_{13} - p_{123}) + (p_{15} - p_{125}) + (p_{23} - p_{123} - p_{234}) \\ & + (p_{24} - p_{234} - p_{245}) + (p_{25} - p_{125} - p_{245}) + (p_{34} - p_{234}) + (p_{45} - p_{245}) \end{aligned}$$

$$P_6(3) = p_{123} + p_{125} + p_{234} + p_{245}$$

$$P_6(x) = 0, \qquad x > 3$$

$$P_6(0) = 1 - P_6(1) - P_6(2) - P_6(3).$$

Let us now find the general relations between the $P_n(x)$ and the p_i, p_{ij}, $\cdots$.

We introduce the sums

$$S_0 = 1, \quad S_1 = \sum_{i=1}^{n} p_i, \quad S_2 = \sum_{i,j}^{1\ldots n} p_{ij}, \quad S_3 = \sum_{i,j,k}^{1\ldots n} p_{ijk}, \ldots, \quad S_n = p_{123\ldots n} \tag{58}$$

and $S_\nu = 0$ for $\nu > n$. Let us express S_1 in terms of the $P_n(x)$. To find S_1, the sum of all the p_i, we add first those regions $R_1, R_2, \ldots, R_n$ which are covered just once; the corresponding probability is $P_n(1)$. Consider next all regions $R_{12}, R_{13}, \ldots$, which are covered exactly twice; the corresponding probability is $P_n(2)$. In computing S_1 the probability corresponding to R_{ij} must be counted twice since R_{ij} belongs both to A_i and A_j; hence, we must add $2P_n(2)$ to $P_n(1)$. Similarly, we must add $3P_n(3)$, where $P_n(3)$ is the probability corresponding to all triply covered regions, etc. The result is

$$S_1 = P_n(1) + 2P_n(2) + \cdots + nP_n(n). \tag{59}$$

To find S_2 we start with $P_n(2)$; to that we have to add $\binom{3}{2}P_n(3)$, corresponding to the regions R_{ijk} [each R_{ijk} belongs to the $\binom{3}{2}$ regions R_{ij}, R_{ik}, R_{jk}] and so on; we find

$$S_2 = P_n(2) + \binom{3}{2} P_n(3) + \binom{4}{2} P_n(4) + \cdots + \binom{n}{2} P_n(n).$$

Continuing in this fashion we obtain the important formulas

$$S_\nu = \sum_{x=\nu}^{n} \binom{x}{\nu} P_n(x), \qquad \nu = 0, 1, ..., n. \tag{60}$$

Formulas (60) are also correct for $\nu = 0$ because of (57), and $S_0 = 1$.

We next introduce the *factorial moments* of $P_n(x)$, defined in Chapter III, Section 4, viz.,

$$m^{(\nu)} = \sum_{x=\nu}^{n} x(x-1) \cdots (x-\nu+1) P_n(x), \qquad \nu = 1, 2, ... n, \quad m^{(0)} = 1. \tag{61}$$

We may use as before the symbol $x^{(\nu)} = \binom{x}{\nu}\nu!$, $x^{(0)} = 1$ where $x^{(\nu)} = 0$ for $\nu > x$ or $\nu < 0$; then $m^{(\nu)} = \sum_{x=\nu}^{n} x^{(\nu)} P_n(x)$. Comparing (60) and (61), we see that

$$S_\nu = \frac{1}{\nu!} m^{(\nu)}, \qquad \nu = 0, 1, 2, ... \tag{62}$$

Thus we see that the sums S_ν are, up to a constant factor, equal to the factorial moments of $P_n(x)$.

We now use (60), for $\nu = 1, 2, ..., n$, as a chain of equations for the "unknowns" $P_n(1), ..., P_n(n)$ [we compute first $P_n(n)$; the next equation contains the single unknown $P_n(n-1)$ etc.] and find easily

$$P_n(x) = \sum_{\nu=x}^{n} (-1)^{\nu+x} \binom{\nu}{x} S_\nu, \qquad x = 0, 1, ..., n. \tag{63}$$

In particular, we find for the probability $1 - P_n(0)$, that at least one of the events takes place, the formula due to H. Poincaré

$$1 - P_n(0) = S_1 - S_2 + S_3 - \cdots + (-1)^{n+1} S_n . \tag{64}$$

These are the basic facts and formulas. For more details we refer to the literature, footnote 1, p. 192. We now turn to some examples.

9. Examples of Arbitrarily Linked Events

9.1. *Jeu de rencontre.* The following simple game can be translated into various different problems. In the "jeu de rencontre" there are in an urn n counters numbered 1 to n. They are drawn out, one at a time, until the urn is empty; there is a "rencontre" if the counter numbered i is drawn at the ith drawing. We ask for *the probability $P_n(x)$ of x rencontres* *in a game* (consisting of n drawings.) Each game consisting of n drawings forms, of course, one element of the collective under consideration. Denote by p_i the probability of "rencontre" at the *i*th drawing, and similarly p_{ij} , ... We have clearly

$$p_i = \frac{1}{n}, \quad p_{ij} = \frac{1}{n(n-1)}, \quad p_{ijk} = \frac{1}{n(n-1)(n-2)}, \ldots, \quad p_{12\ldots n} = \frac{1}{n!} \tag{65}$$

and therefore

$$S_1 = 1, \quad S_2 = \frac{1}{2!}, \ldots, \quad S_\nu = \frac{1}{\nu!}, \ldots . \tag{66}$$

We find the values $P_n(x)$, $x = 1, 2, \ldots, n$ from (63) and $P_n(0)$ from (64):

$$P_n(0) = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} + \cdots \pm \frac{1}{n!}$$

$$P_n(1) = 1 - \frac{2}{2!} + \frac{3}{3!} - \frac{4}{4!} + \cdots = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots \pm \frac{1}{(n-1)!}$$

$$P_n(2) = \frac{1}{2}\left[\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots \pm \frac{1}{(n-2)!}\right]$$

$$\cdots\cdots$$

$$P_n(n-1) = 0$$

$$P_n(n) = \frac{1}{n!} .$$

Thus, for example, for $n = 5 : P_5(0) = 11/30$, $P_5(1) = 3/8$, $P_5(2) = 1/6$, $P_5(3) = 1/12$, $P_5(4) = 0$, $P_5(5) = 1/120$ with sum 1. We see that for an n which is not too small the probabilities $P_n(x)$ decrease markedly with x. The problem is clearly one of "rare events" (Section 5). In fact, the expected value a $(=S_1)$ of $P_n(x)$ equals 1, independent of n.

In this simple case we can easily compute the limit of $P_n(x)$ for fixed x as $n \to \infty$. Compare the expansion of $e^{-1} = (1/2!) - (1/3!) + (1/4!) - \cdots$ with $P_n(1)$, that of $P_n(2)$ with $e^{-1}/2!$, etc. We see that

$$\lim_{n\to\infty} P_n(x) = \frac{1}{ex!} . \tag{67}$$

To the right-hand side of (67) we find the Poisson distribution $\psi(x)$ with $a = 1$, briefly $\psi(x; 1)$

Problem 24. Let

$$\begin{aligned} V_n(x) &= P_n(0) + P_n(1) + \cdots + P_n(x), & x &= 0, 1, \ldots, n \\ &= 1, & x &\geqslant n, \end{aligned}$$

be the probability of the occurrence of at most x events, and

$$\begin{aligned} W_n(x) &= P_n(x+1) + P_n(x+2) + \dots + P_n(n), \qquad & x = 0, 1, \dots, n-1 \\ &= 0, & x \geqslant n, \end{aligned}$$

be the probability of the occurrence of at least $x + 1$ events. Prove that

$$S_\nu = \sum_{x=\nu-1}^{n-1} \binom{x}{\nu - 1} W_n(x), \qquad \nu = 1, 2, \dots, n$$

and

$$W_n(x) = \sum_{\nu=x+1}^{n} (-1)^{\nu+x-1} \binom{\nu - 1}{x} S_\nu \,, \qquad x = 0, 1, \dots, n.$$

Problem 25. Let $f(t) = \Sigma_{x=0}^{n} P_n(x)\, t^x$. Prove that $f(t) = \Sigma_{\nu=0}^{n} S_\nu(t-1)^\nu$.

9.2. *Problem of runs in Bernoulli trials.* We return to the problem of Section 7. As one simplification, we consider the n results to be "cyclic"; by that we mean that we shall also speak of a run of length m when last and first results form together a run of length m; also—more important—we assume as in Section 7.3 that n is large compared to m. We have, in agreement with (44),

$$S_1 = a_m = n(q^m p^2 + p^m q^2).$$

Under the assumption that n/m is very large, we had in (50) $S_2 = \frac{1}{2}n(n-1)(q^m p^2 + p^m q^2)^2$ and one finds in a similar way $S_\nu = \binom{n}{\nu}(q^m p^2 + p^m q^2)^\nu$. These are the same results as in the problem of n repeated trials with $(p^m q^2 + p^2 q^m)$ as the probability of "success." It is therefore easy to determine the asymptotic behavior. We shall return to this problem in Section 9.4.

9.3. *Problem of occupancy.*[1] Let us study a less simple problem. Consider the random placement of k indistinguishable stones in n cells; to each cell corresponds the same probability. If at the end of the placement, there are m stones in a cell we call m the *occupancy number* of the cell.[2] We ask for the probability $P_n^{(m)}(x)$ that among the n cells there are x

[1] See R. v. Mises, "Ueber Aufteilungs- und Besetzungswahrscheinlichkeiten" *Rev. Fac. Sci. Univ. Istanbul* 4 (1938), pp. 145–163.

[2] The cells may be interpreted as cells proper, as days of the week, of the year, as persons, as birthdays, as particles, as accidents, etc.

with m-tuple occupancy. We start with the k-dimensional collective whose elements are the individual arrangements of k indistinguishable figures on n places. There are n^k such arrangements which are considered equally probable. The label space S consists of these n^k "points"; each point has k coordinates taken from the numbers 1, 2, ..., n. We call p_i the probability that the ith cell be one of m-tuple occupancy no matter what happens to the other cells. We show that

$$p_i = \binom{k}{m}\left(\frac{1}{n}\right)^m\left(1-\frac{1}{n}\right)^{k-m}, \qquad i = 1, 2, ..., n. \tag{68}$$

In fact, out of k stones, m stones—for the cell numbered 1—can be selected in $\binom{k}{m}$ ways; the remaining $k - m$ stones may be distributed among the remaining $(n - 1)$ cells in $(n - 1)^{k-m}$ ways, and since n^{-k} is the probability of each arrangement, we find $p_1 = \binom{k}{m}(n - 1)^{k-m}\, n^{-k}$, and this is the right-hand side of (68); the same consideration holds for each of the n places. Let, for example, $k = 3$, $m = 2$, $n = 5$; then the subset A_1 of S, whose probability is p_1, consists of 12 points: each of them has the number 1 for two of the three coordinates and the third coordinate equals 2 or 3 or 4 or 5. Thus, 1, 5, 1 for example, denotes the arrangement where the first and the third stones stand in cell 1 and the second stone in cell 5. We obtain p_1 as $12/5^3$. From (68) and (59) we have

$$S_1 = a_m = \sum_{x=0}^{n} xP_n^{(m)}(x) = p_1 + p_2 + \cdots + p_n = n\binom{k}{m}\left(\frac{1}{n}\right)^m\left(1-\frac{1}{n}\right)^{k-m} \tag{68'}$$

Next we compute p_{12}, the probability that the first and the second cells each be occupied m times, no matter what happens to the other cells. If out of k stones m are put into the first, m into the second cell, and $k - 2m$ stones into the remaining cells, this can be done in $\dfrac{k!}{m!m!(k-2\,m)!}$ ways; this multiplied by $(n - 2)^{k-2m}$ gives the number of all individual arrangements with m stones on each of the first and second cells. Hence, the probability p_{ij} of m-tuple occupancy of the ith and jth cells and arbitrary arrangement in the other $n - 2$ cells equals

$$p_{ij} = \frac{k!(n-2)^{k-2m}}{(m!)^2(k-2m)!}\,n^{-k} = \frac{k!}{m!m!(k-2m)!}\left(\frac{1}{n}\right)^{2m}\left(1-\frac{2}{n}\right)^{k-2m}, \qquad i,j = 1, ..., n. \tag{69}$$

and

$$S_2 = \binom{n}{2}p_{ij}, \qquad m^{(2)} = n(n-1)p_{ij}\,.$$

In the same way we find $p_{12\ldots\nu}$, and from that

$$S_\nu = \binom{n}{\nu} \frac{k!}{(m!)^\nu (k - \nu m)!} \frac{(n - \nu)^{k-\nu m}}{n^k}. \tag{70}$$

From the S_ν we can derive $P_n^{(m)}(x)$ by means of the general formulas (63).

We consider an illustration. Take $n = 365$, $k = 60$, and let a_m as in (68′) be the expected value of the number of cells with occupancy m. We find $a_0 = 309.60$, $a_1 = 51.03$, $a_2 = 4.14$, $a_3 = 0.22$. This may be interpreted as follows. If the birthdays of 60 persons are known, there will be on the average 4 days with double birthdays and 0.22 with triple birthdays; this last result means that among four to five groups of 60 birthdays, there will be on the average one day (in one of these groups) which is a triple birthday. For $n = 365$, $k = 29$, we find $a_2 = 1$; hence from among 29 birthdays, we might on the average find one double birthday.

9.4. *Asymptotic behaviour of the $P_n(x)$.* First, we consider some simple asymptotic results for a_m, given by (68′). If k remains finite while $n \to \infty$, then $a_m \to 0$ for $m \geqslant 2$, while $a_1 \to k$, $a_0 \to \infty$. If k and n both increase in such a way that the average occupancy number $\alpha = k/n$ remains constant, we see from the result of Section 5 that $a_m/n = E_m[x/n]$ approaches the Poisson probability $\psi(m; \alpha)$:

$$\frac{a_m}{n} = E_m \left[\frac{x}{n}\right] \to \frac{e^{-\alpha}\alpha^m}{m!}. \tag{71}$$

If k increases beyond bound while n remains finite, $a_m/n = E_m[x/n]$ converges toward a normal distribution, as we shall verify later.

Can we make a statement regarding the asymptotic behavior of $P_n^{(m)}(x)$ itself, the way we could in the example of Section 9.1 ? There we had explicit and particularly simple expressions for the $P_n(x)$. Here we know so far only S_ν; however, Eqs. (60), which form a chain of equations for the $P_n(x)$, show that knowledge of the S_1, S_2, ..., S_n is equivalent to that of the P_1, P_2, ..., P_n; hence, it should be possible to draw conclusions directly from the S_ν.

Indeed, the so-called *continuity theorem of moments*[3] allows one, under certain circumstances, to conclude that a sequence of functions, whose moments converge toward the moments of a limit function, converges itself towards that limit function. v. Mises[4] has adapted this theorem to

[3] See, e.g., G. Pólya, *Math. Z.* 8 (1920), pp. 171–181; see also Chapter VIII, end of Section 4.

[4] R. v. Mises, "Das Problem der Iterationen." *Z. Angew. Math. Mech.* 1 (1921), pp. 298–307.

the case where the limit distribution is Poisson's function. The resulting criterion is the following. *If,* with the notation (61), *the factorial moments* $m^{(\nu)}$ of $P_n(x)$, $\nu = 1, 2, \ldots,$ *converge toward the* νth *power of a finite constant* as $n \to \infty$

$$\lim_{n\to\infty} m^{(\nu)} = a^\nu, \qquad \nu = 0, 1, 2, \ldots, \tag{72}$$

then the distribution $P_n(x)$ *converges toward the Poisson function with the same "a"*

$$\lim_{n\to\infty} P_n(x) = \psi(x) \equiv \psi(x; a). \tag{73}$$

Let us apply this result to our various examples. We begin with the *problem* of *repeated trials* where the convergence of $P_n(x) \to \psi(x; a)$ has been established directly. In this case we found (Section 4) for the factorial moments $m^{(\nu)} = n(n-1) \cdots (n-\nu+1)\, q^\nu = nq \cdot (n-1)\, q \cdot (n-2)\, q \cdots (n-\nu+1)\, q$, and we see that with $nq \to a$, as in Section 5, $\lim_{n\to\infty} m^{(\nu)} = a^\nu$ as required in (73).

We see also that in the *problem of runs* we have, with the approximation that led to $S_\nu = \binom{n}{\nu}(q^m p^2 + p^m q^2)^\nu$ exactly the same result, namely, $P_n(x) \to \psi(x; a)$, where $a = \lim_{n\to\infty} n(q^m p^2 + p^m q^2)$. In the paper quoted above v. Mises has proved this asymptotic result without neglecting m/n.

In the problem of "*rencontre,*" we had in (66) for all $n : m^{(\nu)} = 1$, $\nu = 1, 2, \ldots$; hence $a = 1$ and there is indeed convergence toward $\psi(x; 1)$.

We now consider in somewhat more detail the *problem of occupancy.* We have seen that if k and n both tend toward infinity in such a way that $k/n \to \alpha$ is finite, then (71) holds. At present, however, we are investigating the asymptotic behavior of $P_n(x)$ and wish a_m and not a_m/n to remain finite. We therefore remember the inequality (35′) (p. 180) which, with the present notation,[5] becomes

$$\left(1 - \frac{1}{n}\right)^{\alpha - m}\left(1 - \frac{m-1}{k}\right)^{m-1} < \frac{E[x/n]}{e^{-\alpha}\alpha^m/m!} < \left(1 - \frac{1}{n}\right)^{-m}. \tag{74}$$

It is seen that $(1 - 1/n)^\alpha = [(1 - 1/n)^n]^{\alpha/n}$ tends to one, even as $\alpha \to \infty$ if $\alpha/n \to 0$. Now $\alpha/n = k/n^2 \to 0$ means that the average occupancy increases less than the number n of cells. We wish to let $\alpha \to \infty$ in such a way that, for m fixed, $ne^{-\alpha}\alpha^m/m!$ tends toward a finite value a as $n \to \infty$; this is achieved if $\alpha \to \infty$ of the order of magnitude of $\log n$ (or more specifically, if $a \sim \log n + m \log\log n$).

[5] For a, n, and x of (35′) we have here α, k, and m, and we use $k/n. \to \alpha$.

Next, with a view to (73), consider the quotient

$$\frac{\nu! S_\nu}{a_m{}^\nu} = \frac{n!}{n^\nu(n-\nu)!}\,\frac{k!(n-\nu)^{k-\nu m}}{(k-\nu m)!(m!)^\nu n^k}\left(\frac{(k-m)!m!n^k}{k!(n-1)^{k-m}}\right)^\nu,$$

where (68′) and (70) have been used. The first fraction approaches 1 for fixed ν and $n \to \infty$. The two other fractions we may write as

$$\frac{k(k-1)\cdots(k-\nu m+1)}{[k(k-1)\cdots(k-m+1)]^\nu}\cdot\left(\frac{n-1}{n-\nu}\right)^{\nu m}\cdot\left(1-\frac{1}{n}\right)^{-\nu k}\left(1-\frac{\nu}{n}\right)^k.$$

The first two factors go toward one, for fixed ν and m, as $k \to \infty$, $n \to \infty$. In considering the last factor, which we call A, we make the above assumption that $\alpha \to \infty$ in such a way that $(\alpha/n) \to 0$. Using the estimate (33), we arrive by simple computations at the inequality

$$\left(1-\frac{\nu}{n}\right)^{\nu\alpha} < A < \left(1-\frac{1}{n}\right)^{-\nu\alpha}.$$

With $(1 - (\nu/n)^{\nu\alpha} = [(1-\nu/n)^{\nu n}]^{\alpha/n}$ we see that the two bounds which enclose A both tend to one even if $\alpha \to \infty$, as long as $\alpha/n \to 0$. Thus, indeed, if $n \to \infty$, $k \to \infty$, $\alpha \to \infty$, $\alpha/n \to 0$:

$$\lim_{n\to\infty}\frac{S_\nu}{a_m{}^\nu} = \frac{1}{\nu!}. \tag{75}$$

Now if we let $\alpha \to \infty$ like $\log n$, then as seen just before, a_m tends toward a finite number "a" and [remember (62)] the $m^{(\nu)}$ converge toward a^ν as in the condition (73). We have the result:

If n, α, and $k = \alpha n$ *tend to infinity in such a manner that* $a_m = n(e^{-\alpha}\alpha^m/m!)$ *tends toward a finite number* a, *then the probability* $P_n^{(m)}(x)$ *for* x *cells with* m*-tuple occupancy converges toward* $\psi(x; a)$ [is approximated by $\psi(x; a_m)$].

All the examples which we have considered so far have the character of rare events, as the reader may verify by considering each time the exact meaning of the respective $P_n(x)$.

It is natural to ask whether and under what conditions there is convergence towards a Gaussian distribution. This question is mathematically much more difficult, as is well known even in the simple case of independent Bernoulli trials. Without entering into a mathematical investigation we wish to explain the character of the problem. If, in the simple problem of Bernoulli trials with mean value $a_n = nq$ and variance $s_n{}^2 = npq$ we let $n \to \infty$ while p and q remain fixed, both a_n and $s_n{}^2$ tend toward infinity. In this case, the Gaussian approximation to the binomial distribution $p_n(x)$ results. We expect to find convergence toward the Gaussian distribution for problems where mean value and variance increase with n. Let us return to the problem of occupancy and consider now the more natural case, *where* $n \rightarrow \infty$, $k \rightarrow \infty$ *while* $\alpha = k/n$ *remains finite*; then the mean value a_m tends to infinity with n while the factor of n in (68′) tends to $(\alpha^m/m!)\, e^{-\alpha} = q$. One can show that in this case $P_n(x)$ *is* indeed *approximated by a Gaussian distribution*[6] *with mean value* nq, *where* $q = (\alpha^m/m!)\, e^{-\alpha}$, α *finite and variance divided by n equal to* $q - q^2[1 + (m - \alpha)^2/\alpha]$.

Problem 26. With $n = 365$, $k = 60$, (as on p. 199) compute for $m = 3$ the probabilities $P_n{}^{(m)}(x)$ for $x = 0$, $x = 1$, $x = 2$ and discuss your result.

D. Application to Mendelian Heredity Theory (Sections 10 and 11)

10. Basic Facts and Definitions

10.1. *Mendel's first law. The basic distributions.* One of the most impressive applications of elementary probability calculus is to the theory of heredity. In the years preceding 1865 Gregor Mendel[1], a Catholic priest and teacher in Austria, investigated in a long series of careful experiments the way in which traits characteristic of certain types of peas are propagated in cross fertilization. In the garden of his monastery he grew generation after generation of pea plants and studied the variations of the form of the seeds, of the shape of the shells, of the color of the flowers, and of several other characters. He discovered that the distribution of these characters, for example that of red and white flowering types among the plants in successive generations, follows certain statistical laws. On the basis of Mendel's experiments and the interpretation suggested by him, there has developed a comprehensive theory, generally accepted today by geneticists, the simplest case of which will be discussed in this section.

Assume that a definite quality, a character, or "Mendelian character"

[6] Geiringer, *loc. cit.* p. 192 (see p. 22 of the paper).

[1] The first publication "Versuche über Pflanzenhybriden" appeared in the *Verhandlungen des Naturforschenden Vereins in Bruenn. IV. Band, Abhandlungen, 1865*; Bruenn, 1866, pp. 3-47. English translation: "Experiments on Plant Hybridization " Harvard Univ. Press, Cambridge, 1946.

The deep going recent molecular biological findings do not invalidate the statistical regularities we want to discuss.

of a plant can take on several, say r, distinct values. The color of the flower of peas is such a Mendelian character, where $r = 2$, and the alternatives are red and white. According to the Mendelian theory, each individual plant, with respect to this character, is biologically characterized by *two* of the r values, called factors or genes. They determine the biotype or *genotype* of the plant. In our example the biotype may thus be red-red, or white-red, or white-white. It does not interest us here how the actual appearance (phenotype) of the individual depends on the two factors, which of them is "dominant," etc. In the process of bisexual reproduction, each parent individual transmits one of its two color-genes to the descendant, so that a new individual with two color-genes comes into existence. More specifically, each parent segregates and transmits a gamete (egg or sperm); with respect to each character, the gamete contains then one gene which has been selected from among the two genes related to this character, contained in the parent cell. The assumption is made that for an individual with genes a_i, a_j, the probability of segregating and transmitting a gamete with either a_i or a_j is $\frac{1}{2}$ and that the union of the two gametes (sperm and egg) again happens "at random," a term whose meaning in this connection will be explained in more detail.

In the preceding paragraph we explained the essential content of *Mendel's first law.* Let us give a more mathematical formulation. We assume that, corresponding to each character there is a random variable, x, which takes on r distinct values $a_1, a_2, \dots, a_r$, the r *alleles.* In the example of color of peas, r equals two; a_1 stands for "red," a_2 for "white." Three alleles determine the human blood groups and there are at least fifteen, probably many more, alleles for the eye color of drosophila. The genetic type of an organism is specified by two values x and y of this random variable equal to each other or not; they represent the organism's maternal and paternal heritage. We assume that two individuals are genotypically the same with respect to a given character even if in one of them x comes from the individual's mother, y from its father, while in the other it is the other way around. If we denote the type of an organism by $(x; y)$, where x denotes the maternal, y the paternal heritage then $(x; y) = (y; x)$ and therefore, in the simple case which we are discussing here, we do not need the semicolon. Consequently, there are $r(r + 1)/2$ possible genotypes, and we assume that in a certain (initial) generation these types are distributed according to a certain probability law, the *distribution of genotypes*, $w^{(0)}(x, y)$, which we assume to be the same for males and females. From $w^{(0)}(x, y)$, the distribution in the nth generation must follow by means of the rules given above. We add the assumption that there are distinct, non-overlapping generations.

Denote now the distribution in the nth generation by $w^{(n)}(x, y)$:

$$w^{(n)}(x,y) = w^{(n)}(y,x), \qquad n = 0, 1, \ldots, \qquad \begin{matrix} x \\ y \end{matrix} = a_1, a_2, \ldots, a_r \tag{76}$$

and

$$\sum_x \sum_y w^{(n)}(x,y) = 1. \tag{76'}$$

Next denote by l_0 (by l_1) the probability that in the process of segregation a parent transmits its paternal (maternal) gene and we assume that, in this simplest case,

$$l_0 = l_1, \qquad l_0 + l_1 = 1; \tag{77}$$

hence, each $l_i = \frac{1}{2}$, as stated earlier. The *segregation distribution*, given here by (77) is much less trivial in other situations. From these two distributions (76) and (77) we derive the third basic distribution, the *distribution of gametes* $p^{(n)}(x)$, $n = 0, 1, \ldots$; $x = a_1, a_2, \ldots, a_r$; this is the probability that in the nth generation a gamete possesses the gene x (with respect to the character in question).

Let us compute $p^{(n)}(x)$: in order to transmit the gene x the parent must posses it, and transmit it. Hence we have, e.g., for $x = 1$ (writing now 1, 2, ..., r for $a_1, a_2, \ldots, a_r$), $p^{(n)}(1) = w^{(n)}(1, 1)(l_0 + l_1) + w^{(n)}(1, 2)l_1 + w^{(n)}(2, 1)l_0 + \cdots + w^{(n)}(r, 1)l_0$. For example, the term $w^{(n)}(1, 2)l_1$ is the probability that the parent possesses the gene 1 as maternal gene, the gene 2 as paternal gene, and transmits its maternal gene. The sum of all such terms gives the probability of a gamete of type 1. Because of Eqs. (76) and (77), the above may be written as

$$p^{(n)}(x) = \sum_y w^{(n)}(x,y), \qquad x = 1, 2, \ldots r; \qquad n = 0, 1, \ldots \tag{78}$$

with

$$\sum_{x=1}^{r} p^{(n)}(x) = 1.$$

Under *random mating* (see later) a new individual of the $(n + 1)$th generation is then formed by the fusion of two gametes of the nth generation. The assumption that this fusion happens "at random" is expressed by

$$w^{(n+1)}(x,y) = p^{(n)}(x)p^{(n)}(y), \qquad \begin{matrix} x \\ y \end{matrix} = 1, 2, \ldots, r \tag{79}$$

and that finishes the cycle since we now know $w^{(n+1)}(x, y)$, the distribution of genotypes in the next generation.

10.2. *Constancy of gametic proportions.* It is now very easy to derive the famous law of constancy of gametic proportions, recognized by the biologist W. Weinberg,[2] proved by G. H. Hardy,[3] and verified by a very great number of observations. We obtain from (78) and (79), for $n = 0, 1, \ldots,$

$$p^{(n+1)}(x) = \sum_{y=1}^{r} w^{(n+1)}(x, y) = \sum_{y=1}^{r} p^{(n)}(x)p^{(n)}(y) = p^{(n)}(x)\sum_{y=1}^{r} p^{(n)}(y) = p^{(n)}(x). \tag{80}$$

This line contains the complete proof. From (80) follows

$$w^{(n+1)}(x, y) = w^{(n)}(x, y), \qquad n = 1, 2, \ldots . \tag{80'}$$

This basic result states that under random mating and for one single character *the distribution of gametes remains constant throughout while that of the genotypes remains constant from the first filial generation on.*

As an example, consider the character "color," let $r = 2$, that is, only two colors are in question. There are three genotypes: 1, 1; 1, 2; 2, 2 (white-white, white-red, red-red). The original distribution may be $w^{(0)}(1, 1) = 0.40$, $w^{(0)}(1, 2) = w^{(0)}(2, 1) = 0.25$, $w^{(0)}(2, 2) = 0.10$; that is, 40% of all plants are pure-race white, 10% pure-race red, and 50% are hybrids. The scheme of p-values is $p^{(0)}(1) = 0.40 + 0.25 = 0.65$, $p^{(0)}(2) = 0.35$, and we obtain $w^{(1)}(1, 1) = (0.65)^2 = 0.4225$, $w^{(1)}(1, 2) = w^{(1)}(2, 1) = (0.65)(0.35) = 0.2275$, $w^{(1)}(2, 2) = 0.1225$. If we next compute $p^{(1)}(x)$, we find indeed $p^{(1)}(1) = 0.4225 + 0.2275 = 0.65$, $p^{(1)}(2) = 0.35$, and consequently $w^{(2)}(x, y) = w^{(1)}(x, y)$, and this distribution will remain. Thus in the first filial generation there will be 42% of type white-white, 12% of type red-red, and 46% hybrids. If we start with plants of pure race in the first generation, having $w^{(0)}(1, 1) = w^{(0)}(2, 2) = 0.50$, the same computation supplies $w^{(1)}(1, 1) = 0.25$, $w^{(1)}(2, 2) = 0.25$, $w^{(1)}(1, 2) = w^{(1)}(2, 1) = 0.25$; with red "dominant" this implies that from equal numbers of white and red plants of pure race, we get $\frac{3}{4}$ red and $\frac{1}{4}$ white in the second generation. This and other results have been confirmed by extensive experiments.

10.3. *On random mating.* Let us return to the concept of *random mating*, which is an essential hypothesis in the preceding considerations. In the

[2] W. WEINBERG, "Über Vererbungsgesetze beim Menschen " *Z. Induktive Abstammungs u. Vererbungslehre* 1 (1909), pp. 277–330.

[3] G. H. HARDY, "Mendelian proportions in a mixed population " *Science* **28** (1908), pp. 49–50. Before Hardy, K. Pearson, in 1903, proved a special case.

simplest case, discussed above, the mathematical meaning of random mating is contained in Eqs. (78) and (79), by means of which $w^{(n+1)}$ is derived from $w^{(n)}$. In an equation of type (78) the distribution of gametes $p^{(n)}(x)$ appears as a linear expression in the $w^{(n)}(x, y)$ with constant coefficients, which are sums of segregation probabilities, while Eq. (79) expresses the random fusion of two gametes. In the various cases of selection, one or the other of the assumptions which lead to (78) and (79) is no longer made.

Problem 27. The 3 pure races that correspond to a three-valued Mendelian character are distributed in the first generation in the ratio 1 : 4 : 5. Compute the distribution in the second generation and check Hardy's theorem by computing the third generation.

Problem 28. A two-valued Mendelian character is transmitted from the first to the second generation according to Mendel's theory. Assume $w^{(0)}(1, 1) = w^{(0)}(2, 2) = w^{(0)}(1, 2) = \frac{1}{4}$. The three genotypes AA, AB, BB have different survival factors λ, μ, $\nu < 1$; that is, out of n individuals of the type AA, only λn survive to reach the reproduction age. Compute the distribution of genotypes in the second generation and discuss the result for various λ, μ, ν in relation to Hardy's theorem.

Problem 29. Assume that there is a "rule" that only identical types may mate. Let $r = 2$. Prove that in this case

$$w^{(n+1)}(1, 1) = w^{(n)}(1, 1) + \tfrac{1}{2} w^{(n)}(1, 2) .$$

Find $w^{(n+1)}(1, 2)$ and $w^{(n+1)}(2, 2)$.

11. Probability Theory of Linkage[1]

11.1. *Mendel's second law. Linkage distribution.* We return to random mating and to the same problem as in the first part of Section 10, but now we consider simultaneously several, say m, characters of an individual, e.g., color of the flower, shape of the seed, length of the stem, etc. This rather than one single character is, of course, the basic situation of genetics. A genotype is now described by two sets of m numbers each, $(x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_m) \equiv (x; y)$; here the letters before (after) the semicolon designate the individual's maternal (paternal) heritage such that x_i (or y_i) is the gene related to the ith character inherited from the individual's mother (father), $i = 1, 2, \ldots, m$. The distinction between maternal and paternal heritage is essential for the understanding of

[1] As the term linkage indicates, "dependence" plays the essential role in this problem.

linkage: a genotype is characterized not by $2m$ numbers but by two m-dimensional vectors. Thus, with respect to inheritance, the types $(1, 2; 5, 7)$ and $(1, 7; 5, 2)$ are different although both possess the genes 1 and 5 with respect to the first and the genes 2 and 7 with respect to the second character. On the other hand, $(x; y) = (y; x)$, i.e., $(1, 2; 5, 7) = (5, 7; 1, 2)$.

A *gamete* contains m genes, one with respect to each character. The kinds of gametes which an organism may produce depend on the combination of the genetic material it has inherited. Clearly there exist 2^m such possible combinations. Mendel's original assumption was that all these combinations have equal probability. This is *Mendel's second law* or the *law of independent assortment*[2]; in this case 2^{-m} is the probability for each of the possible gametes.

This simple conception was shaken by the observation of "linkage." The deviation from independent assortment, called linkage, was first pronounced for $m = 2$ when it was observed that for the organism $(x_1, x_2; y_1, y_2)$ the gametic combinations (x_1, x_2) and (y_1, y_2) appear more frequently (have greater probability) than (x_1, y_2) and (y_1, x_2), viz., with probabilities $(1 - c)/2$ and $c/2$, respectively, where $c < \frac{1}{2}$; in other words, the idea is that the genes that came together exhibit a tendency to stay together. In the general case of m characters, there are $m(m - 1)/2$ "crossover values" c_{ij}, where c_{ij} is the probability that either x_i and y_j be transmitted or y_i and x_j, no matter what happens to the $(m - 2)$ other factors. (For $m = 2$ we put $c_{12} = c$) It will, however, be seen that for $m > 3$ the c_{ij} do not suffice to characterize the segregation possibilities. Hence, we introduce a general *linkage distribution*, l.d., as follows. Denote by S the set of m numbers 1, 2, ..., m, by T any subset of S $(0 \subseteqq T \subseteqq S)$, and by l_T *the probability that the maternal genes belonging to T and the paternal genes belonging to $T' = S - T$ be transmitted.* (This definition is independent of the genotype of the parent.) There are 2^m such probabilities with sum one. For symmetry reasons we assume

$$l_T = l_{T'}, \tag{81}$$

and hence, there remain $2^{m-1} - 1 = M$ probabilities. For $m = 2$, $M = 1$, for $m = 3$, $M = 3$, and hence, in these cases the number M of

[2] The case of *independent assortment* has been completely investigated by H TIETZE, "Über das Schicksal gemischter Populationen nach den Mendelschen Vererbungsgesetzen." *Z. Angew. Math. Mech.* 3 (1923), pp. 362–393. The general problem of this section has been solved completely by H. GEIRINGER, "On the probability theory of linkage in Mendelian heredity." *Ann. Math. Statist.* 15 (1944), pp. 25–57. In this paper and in H. GEIRINGER, "On some mathematical problems arising in Mendelian genetics." *J. Amer. Statist. Assoc.* 44 (1949), pp. 526–547, relevant bibliography may be found.

parameters l equals that of crossover values, but for $m > 3$ there are more l-values than crossover values. The linkage distribution may be written as an m-dimensional alternative $l(\epsilon_1 , \epsilon_2 , ..., \epsilon_m)$, with $\epsilon_i = 0$ or 1, where 0 denotes the paternal, 1 the maternal gene. Thus, for example, for $m = 6$, $l(1, 1, 1, 0, 0, 1)$ is the probability that the genes segregated by an organism $(x; y)$ are $x_1 , x_2 , x_3 , y_4 , y_5 , x_6$. We denote by $p^{(n)}(z_1 , ..., z_m)$ the distribution of gametes in the nth generation and by $w^{(n)}(x_1 , ..., x_m ; y_1 , ..., y_m)$ the distribution of genotypes in the nth generation.

11.2. *Recursion problem*. The main problem is that of recursive relations corresponding to relations (78)–(80). We consider it for the $p^{(n)}(z_1 , z_2 , ..., z_m)$ since this is simpler than for $w^{(n)}(x_1 , x_2 , ..., x_m ; y_1 , y_2 , ..., y_m)$. We remember the concept of *marginal distribution* introduced in Chapter III. For example, $\Sigma_{z_3} \Sigma_{z_4} \cdots \Sigma_{z_m} p(z_1 , z_2 , z_3 , z_4 , ..., z_m) = p_{12}(z_1 , z_2)$ is the probability of the result z_1 , z_2. We write $p_T(z_T)$ for $p_{ij \cdots k}(z_i , z_j , ..., z_k)$ if T denotes the set $(i, j, ..., k)$ and z_T the set $(z_i , z_j , ..., z_k)$. If finally, we write $p^{(n)}(z)$ for $p^{(n)}(z_1 , z_2 , ..., z_m)$, our fundamental recursive formula, which we here quote without proof (see the paper by Geiringer, *loc. cit.* p. 207), is

$$p^{(n+1)}(z) = \sum_{(T)} l_T p_T(z_T) p_{T'}(z_{T'}), \tag{82}$$

where the summation is over all subsets T of S. We have, for example,

$$m = 2: \quad p^{(n+1)}(z_1 , z_2) = 2l(00)p^{(n)}(z_1 , z_2) + 2l(01)p^{(n)}(z_1)p^{(n)}(z_2)$$

$$m = 4: \quad p^{(n+1)}(z_1 , z_2 , z_3 , z_4) = 2\{l(0000)p^{(n)}(z_1 , z_2 , z_3 , z_4)$$
$$+ [l(1000)p_1^{(n)}(z_1)p_{234}^{(n)}(z_2 , z_3 , z_4) + \cdot + \cdot + l(0001)p_4^{(n)}(z_4)p_{123}^{(n)}(z_1 , z_2 , z_3)]$$
$$+ [l(1100)p_{12}^{(n)}(z_1 , z_2)p_{34}^{(n)}(z_3 , z_4) + \cdot + \cdot]\}.$$

It is not difficult to interpret this formula. Note that $2l(1000) = l(1000) + l(0111)$, etc., and consider, for example, the term $l(1000)p_1^{(n)}(z_1)p_{234}^{(n)}(z_2 , z_3 , z_4)$. If we denote by x a maternal, by y a paternal heritage, then a typical term of the sum which constitutes the above product is $l(1000)p^{(n)}(z_1 , x_2 , x_3 , x_4) \cdot p^{(n)}(y_1 , z_2 , z_3 , z_4) = l(1000)w^{(n+1)}(z_1 , x_2 , x_3 , x_4 ; y_1 , z_2 , z_3 , z_4)$; this product is the probability that a genotype which has z_1 as its first maternal character, z_2 , z_3 , z_4 , as its second, third, and fourth paternal characters, segregates (in the formation of a gamete) its first maternal and its second, third, and fourth paternal characters, thus producing a gamete of type

(z_1, z_2, z_3, z_4). We recognize that the clarity of the basic formula (82) is due to the linkage distributions whose values act as separators.

The next task consists of the *solution of the recursive equations* (82), which express $p^{(n+1)}(z_1, z_2, ..., z_m)$ in terms of the $p^{(n)}(z_1, z_2, ..., z_m)$, and of all marginal distributions of $p^{(n)}$. This is a system of quadratic difference equations, from which we eventually wish to determine the $p^{(n+1)}$ in terms of the $p^{(0)}$. For the solution of this rather complicated problem see the paper quoted above. Here we note only a few more immediate results: by applying an $(m - 1)$-tuple summation to both sides of (82) we have, for a single gene z_i:

$$p_i^{(n)}(z_i) = p_i^{(0)}(z_i), \tag{83}$$

where the p_i are marginal distributions of first order.[3]

11.3. *Asymptotic behavior.* It is not difficult to study the limiting behavior of $p^{(n)}(z_1, z_2, ..., z_m)$ as $n \to \infty$. Consider first the simple case of *complete linkage*, where $l(00 \cdots 0) = l(11 \cdots 1) = \frac{1}{2}$; this means that non-vanishing segregation probabilities exist only in the event that either *all* maternal genes, or *all* paternal genes are transmitted together. In this case all m genes act like a single one and the recursive formula shows immediately that

$$p^{(n)}(z_1, z_2, ..., z_m) = p^{(0)}(z_1, z_2, ..., z_m), \quad \text{complete linkage;} \tag{83$'$}$$

hence, there is equilibrium for any n. If $m > 1$ this is the only case of equilibrium for finite n. If *all* $c_{ij} > 0$, one can prove that

$$\lim_{n\to\infty} p^{(n)}(z_1, z_2, ..., z_m) = p_1^{(0)}(z_1)p_2^{(0)}(z_2) \cdots p_m^{(0)}(z_m); \tag{84}$$

this means: *the distribution of gametes converges toward independent assortment.* The $p_1^{(0)}(z_i)$ are the marginal distributions of first order of the original distribution. A corresponding result holds in the limit for the distribution of genotypes. Finally, if $c_{ij} > 0$ does not hold for *all* pairs i, j, the linkage distribution degenerates in various ways into $t < m$ groups of complete linkage. Assume, e.g., $m = 8$, and that the groups (1, 4) as well as (2, 3, 5, 6) are completely linked (i.e., $c_{14} = 0$, $c_{23} = c_{25} = \cdots = c_{56} = 0$) while the values c_{78}, c_{12}, c_{17}, c_{18}, c_{27}, and c_{28} are known to be different from zero; that means there is at least one crossover possibility between any two of the four groups (1, 4), (2, 3, 5, 6), (7), (8); then, for all n, $p_{14}^{(n)} = p_{14}^{(0)}$, $p_{2356}^{(n)} = p_{2356}^{(0)}$, and

[3] This is the mathematical expression of the "immortality" of the genes, the generalization of Hardy's law.

$\lim_{n\to\infty} p^{(n)}_{12\cdots8} = p^{(0)}_{14}p^{(0)}_{2356}p^{(0)}_{7}p^{(0)}_{8}$. The generalization we can draw from this example is obvious.

We have discussed in this section a general and important problem of heredity which is at the same time an interesting problem of elementary probability calculus. Very many probability problems appear in genetics and some of them are amenable to a very satisfactory mathematical treatment.

E. Comments On Markov Chains (Sections 12 and 13)

12. Definitions. Classification

12.1. *Definition of a Markov chain. Transition probabilities.* We have considered in this chapter varied examples of dependent trials. The theory of arbitrarily linked events deals with a *general* dependence, but for *two* attributes. One starts with n alternatives, as in the Bernoulli problem, but the probability of success in the ith and jth trial is p_{ij} rather than p^2, etc. From these data we then derive $P_n(x)$, $x = 0, 1, ..., n$, the probability of x events in n trials, in analogy to the Bernoulli distribution $p_n(x)$. In Mendelian heredity the linkage distribution is a general m-dimensional probability, but again, each ϵ_μ can take on two values only.

The so-called Markov chains deal with what is perhaps the most immediate generalization of independent trials, *a particular form of dependence*. Here, the original variable x can take on any finite number r of values, or countably many, or it may even be continuous. Markov chains[1] have been extensively and systematically investigated and are of great interest, both theoretically and from the point of view of applications. We present here a few explanations and examples concerning *finite* Markov chains with *constant* transition probabilities. Not all proofs will be carried out.

We explain in terms of a scheme of urns. Consider r urns numbered $1, 2, ..., r$. Each urn contains the numbers $1, 2, ..., r$ or $A_1, A_2, ..., A_r$ in

[1] A. A. Markoff, *Wahrscheinlichkeitsrechnung*. Leipzig, Berlin, 1912 (no English translation). The "Anhang II," p. 272 is a translation from A. A. Markov, "Extension des théorèmes limites du calcul des probabilités à la somme des valeurs liées en chaîne." *Mem. Acad. Imp. Sci. St. Petersbourg, Cl. Phys. Math.* [8] 22 (1908). The papers by Markov on the present subject appeared between 1907 and 1924; they are quoted with great completeness in the monograph of M. B. Hostinsky, *Méthodes générales du calcul des probabilités*. Mémorial des Sciences Math. LII, 1931, pp. 60/63. This book contains also a presentation of Hostinsky's own contributions to the subject.

some composition. In addition to these r urns there is one "initial" urn, denoted by zero, which contains the r numbers in the *initial distribution* $p_i^{(0)}$, $i = 1, 2, ..., r$. If A_j is the result of the initial drawing from this urn, the urn designated by "j" is used for the next drawing. Its distribution is p_{ij} , $i = 1, 2, ..., r$, where the second subscript denotes the number of the urn and the first, the possible results. If, next, A_k is the result of the drawing from this urn, we turn to urn "k" for which p_{ik} is the probability of the result i, and so on. Thus, p_{ij} is the *conditional probability of drawing the result i from the urn numbered j.* (Since the p_{ij} are conditional probabilities, p_{ij} being the probability of the result i if the preceding result was j, we follow the usual notation for conditional probabilities.)

The characteristic property of such a *Markov chain of first order* is that the probability for the $(m + 1)$th trial depends upon the result of the mth trial but is independent of prior outcomes. In a *chain of order s* the probability of the $(m + s)$th trial depends on the results of the s immediately preceding trials but is independent of prior outcomes. It is easily shown that a chain of order s can be formally reduced to a chain of order 1.

In a chain of order one our definition of the p_{ij} reads

$$p_{ii} = \Pr\{x_{m+1} = i \mid x_m = j\} = \frac{\Pr\{x_{m+1} = i, x_m = j\}}{\Pr\{x_m = j\}},$$

with the assumption that the denominator-probability is greater than zero. If we want to emphasize that the p_{ij} are independent of m, the serial number of the trial in question, we speak of *stationary transition probabilities.* (An important generalization concerns transition probabilities changing with m.) We use the term Markov chain for the scheme just described and also for the sequence of numbers so obtained. It has to be understood that *this sequence of numbers does not form a collective.* There is indeed no randomness in the sequence of outcomes since the probability of a label depends on the preceding result.

From the above definitions it follows that, $i, j = 1, 2, ..., r$

$$p_i^{(0)} \geqslant 0, \qquad p_{ij} \geqslant 0, \qquad \sum_{i=1}^{r} p_i^{(0)} = 1, \qquad \sum_{i=1}^{r} p_{ij} = 1. \tag{85}$$

Another interpretation of a Markov chain is in terms of *a system S* which can be in any of r possible states A_1 , A_2 , ..., A_r . At given equidistant discrete moments t_1 , t_2 , ... the system changes state. The probability that S is in a state A_i at time t_{m+1} depends on the state of S

at time t_m, and not on m or on the state at any time prior to t_m: p_{ij} is the probability that S reaches the state A_i in one transition, starting from A_j. (If the t_m are not *given* instants, but instead time is considered a random variable we speak of a stochastic process.[2])

The so-called *state diagram* is very suggestive (Fig. 12). Consider, for example, three states A_1, A_2, A_3 and the transition probabilities $p_{11} = p_{22} = p_{13} = 0$, $p_{12} = p_{21} = \frac{1}{4}$, $p_{23} = \frac{1}{8}$, $p_{31} = p_{32} = \frac{3}{4}$, $p_{33} = \frac{7}{8}$. In Fig. 12 the arrows represent the transition probabilities. If, at any time, the system is in state A_1 the figure shows that it has probability $\frac{3}{4}$ of reaching A_3 and probability $\frac{1}{4}$ of reaching A_2. If then at some instant it is in A_3 it has probability $\frac{7}{8}$ of staying there and probability $\frac{1}{8}$ for a transition to A_2 while the probability from A_3 to A_1 is zero, and accordingly no arrow points from A_3 to A_1.

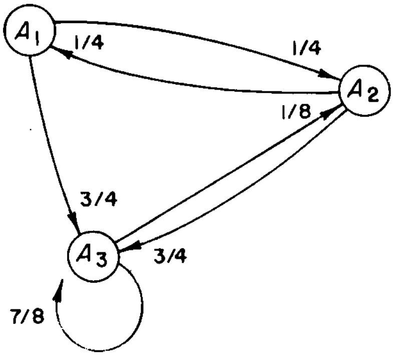

FIG. 12. A state diagram.

The r discrete r-valued distributions with elements p_{ij} may be arranged in an $(r \times r)$-matrix the *transition matrix*, P

$$P = \begin{pmatrix} p_{11} & p_{12} \cdots p_{1r} \\ p_{21} & p_{22} \cdots p_{2r} \\ \vdots & \\ p_{r1} & p_{r2} \cdots p_{rr} \end{pmatrix}. \tag{86}$$

In the ith row are the probabilities leading *to* the state A_i, in the jth column those *from* state A_j; all elements of this square matrix are $\geqslant 0$ and *the sum of each column equals one*. Such a matrix is called *a stochastic matrix*.

[2] See a few comments in Appendix Four.

12.2. *Absolute probabilities and higher transition probabilities.* We call $p_i^{(m)}$ the probability that after m transitions the system reaches the state A_i. The $p_i^{(m)}$ are absolute probabilities in contrast to the conditional probabilities p_{ij}, $i, j = 1, 2, ..., r$. We wish now to find the $p_i^{(m)}$, given the distribution $p_i^{(0)}$ and the p_{ij}. Clearly the following recursion formula holds:

$$p_i^{(m)} = \sum_{j=1}^{r} p_{ij} p_j^{(m-1)}, \qquad i = 1, 2, ..., r, \qquad m = 1, 2, \dots . \tag{87}$$

Thus, if we know the initial distribution $p_i^{(0)}$ and the matrix P we can indeed compute $p_i^{(m)}$. Summing Eqs.(87) from $i = 1$ to $i = r$ we see that

$$\sum_{i=1}^{r} p_i^{(m)} = \sum_{j=1}^{r} p_j^{(m-1)} \sum_{i=1}^{r} p_{ij} = \sum_{j=1}^{r} p_j^{(m-1)} = \cdots = \sum_{i=1}^{r} p_i^{(0)} = 1.$$

We may consider the r numbers $p_i^{(m)}$ as the components of a column vector $\mathbf{p}^{(m)}$; then Eq. (87) shows that $\mathbf{p}^{(m)}$ is derived from $\mathbf{p}^{(m-1)}$ by multiplication with the matrix P. Therefore

$$\mathbf{p}^{(m)} = P\mathbf{p}^{(m-1)} = P^2\mathbf{p}^{(m-2)} = \cdots = P^m\mathbf{p}^{(0)}. \tag{87'}$$

The behavior of $\mathbf{p}^{(m)}$, in particular as $m \to \infty$, depends therefore on that of P^m.

The study of stationary chains with r states was started by Markov (see p. 210) who obtained basic results. Important early contributions —on finite Markov chains—are due to B. Hostinsky, R. v. Mises, M. Fréchet, V. Romanovsky, and G. Frobenius.[3]

From $p_{ik}^{(1)} = p_{ik}$ we can compute probabilities $p_{ik}^{(m)}$ *for a transition from A_k to A_i in exactly m steps.* Clearly:

$$p_{ik}^{(2)} = \sum_{j=1}^{r} p_{ij} p_{jk}, \qquad \text{and} \qquad \sum_{i=1}^{r} p_{ik}^{(2)} = 1, \qquad i, k = 1, 2, ..., r.$$

The right-hand side exhibits the rule of matrix multiplication and if we denote the matrix whose general term is $p_{ik}^{(2)}$ by P_2, we have $P_2 = P^2$. In general

$$P_m = P^m, \qquad m = 1, 2, \dots, \tag{88}$$

[3] See v. Mises [21], p. 532 ff. The monograph of Hostinsky (see p. 210) is also from 1931. See Fréchet [9]. Important earlier papers are by Fréchet, Romanovsky, and in particular by Frobenius. Modern textbook presentations, some going far beyond the simple case presented here, are by Doob [6], Blank-Lapierre and Fortet [2], Feller [7b], Chapters 14 and 15, and the monograph, Chung [3], of 1960.

where P_m is the matrix of the $p_{ik}^{(m)}$. For $m > n$

$$P^m = P^n \cdot P^{m-n},$$

or

$$p_{ik}^{(m)} = \sum_{j=1}^{r} p_{ij}^{(n)} p_{jk}^{(m-n)}, \qquad n = 1, 2, ..., m - 1. \tag{89}$$

Equation (89) is the simplest instance of the *Chapman-Kolmogorov equation.*

We have seen in Eqs. (88) and (87′) that the $p_{ij}^{(m)}$ depend on P^m and the $p_i^{(m)}$ on P^m and the initial distribution. *Hence the limit behavior of the chain depends on P^m.* Essential methods of investigation are based on the *structural properties* of P, or, roughly speaking, on properties which depend on the location of the zeros in P. Also the *characteristic roots* of P, i.e., with I the unit matrix, the roots λ of the determinental equation $| P - \lambda I | = 0$ form an important tool of characterization, due to the fact that the characteristic roots of P^m equal λ^m.

12.3. *First remarks on classification.* Following Hadamard and Fréchet we call a Markov chain *regular* if $p_i^{(m)}$ converges toward a limit distribution u_i, i.e., if $p_i^{(m)} \to u_i$ independent of the initial distribution $p_i^{(0)}$, or, equivalently, if $p_{ik}^{(m)} \to u_i$ for all k. If all $u_i > 0$ the chain is called *positive regular* and this is actually the most important case. Markov himself *proved that $p_i^{(m)} \to u_i > 0$ if all $p_{ij} > 0$ (sufficient condition).* If a limit of the $p_i^{(m)}$ exists but depends on the initial distribution, or, equivalently, if $p_{ik}^{(m)} \to u_{ik}$ we call the chain *semi-regular.* If no limit exists, the term *singular* is often used (this term has nothing to do with vanishing of the determinant of P). We shall see that in this last case the $p_i^{(m)}$ are asymptotically periodic.

Equations (87) have the form of an iteration setup for solving the system of homogeneous equations

$$-x_i + \sum_{j=1}^{r} p_{ij} x_j = 0, \qquad i = 1, 2, ..., r. \tag{90}$$

The determinant of this homogeneous system vanishes, as seen by considering the transposed system

$$-y_i + \sum_{j} p_{ji} y_j = 0, \qquad i = 1, 2, ..., r \tag{91}$$

which, on account of the last Eq. (85) has the non-vanishing solution $y_i = 1, i = 1, 2, ..., r$. Thus (90) has a non-zero solution $x_i = u_i$, which we call the *stationary solution*, since for $p_i^{(0)} = u_i$, the $p_i^{(m)}$ will remain the same for all m. Clearly, a unique stationary solution may exist without the $p_i^{(m)}$ converging toward it. We shall have to investigate under what conditions the $p_i^{(m)}$ actually converge toward the u_i for arbitrary $p_i^{(0)}$.

The fact that the determinant of Eq. (90) vanishes is equivalent to the statement that $\lambda = 1$ *is a root of the characteristic equation* which is obtained by equating to zero the determinant $P(\lambda)$ of the equations

$$\sum_{j=1}^{r} p_{ij}x_j - \lambda x_i = 0, \qquad i = 1, 2, ..., r. \tag{92}$$

We have

$$P(\lambda) \equiv |P - \lambda I| = \begin{vmatrix} p_{11} - \lambda & p_{12} & \cdots p_{1r} \\ p_{21} & p_{22} - \lambda & \cdots p_{2r} \\ \vdots & & \\ p_{r1} & p_{r2} & \cdots p_{rr} - \lambda \end{vmatrix} = 0. \tag{93}$$

It can be shown very easily[4] that *no root of* (93) *can be greater than one in absolute value.*

We consider now a few illustrative examples:

(1) For a *positive regular case*, take $r = 3$, $p_{11} = p_{22} = p_{33} = 0$, all other terms equal to $\frac{1}{2}$. We solve the equations

$$u_1 = \tfrac{1}{2}u_2 + \tfrac{1}{2}u_3, \qquad u_2 = \tfrac{1}{2}u_1 + \tfrac{1}{2}u_3$$

and find $u_1 = u_2 = u_3 = \frac{1}{3}$. Computation of P^n or of $\mathbf{p}^{(n)}$ shows geometric convergence toward these values. The characteristic roots are $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = -\frac{1}{2}$.

(2) Let $r = 3$ and $p_{11} = \frac{7}{8}$, $p_{22} = \frac{5}{8}$, $p_{33} = 1$, $p_{12} = \frac{3}{8}$, $p_{21} = \frac{1}{8}$; all other terms are zero. (The reader should each time set up the matrix P and the state diagram.) Here, the u_i are not uniquely determined; one obtains (c being a constant): $u_1 : u_2 : u_3 = 3 : 1 : c$, or, in other words, the stationary solution depends on $\mathbf{p}^{(0)}$. This is an example of the *semi-regular case.* The simplest example of that case is given for $r = 2$, by $p_{11} = p_{22} = 1$, $p_{12} = p_{21} = 0$ with $u_1 = c$, $u_2 = 1 - c$. In both these examples the matrix P is of the type $P = \left(\begin{smallmatrix} P_1 & 0 \\ 0 & P_2 \end{smallmatrix}\right)$, where P_1 and P_2 are square matrices of orders r_1 and $r_2 = r - r_1$ and all other terms of P are zero. Such a matrix is called *completely reducible*. There are two distinct groups of states which *remain* distinct.

[4] Consider $|P^m - \lambda^m I| = 0$.

In the first example the roots are $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = \frac{1}{2}$, in the second $\lambda_1 = \lambda_2 = 1$. It is easily understood why $\lambda = 1$ is a multiple root of $P(\lambda) = 0$: in fact, each of the matrices P_1, P_2 is a stochastic matrix and has $\lambda = 1$ as characteristic root, and $P(\lambda) = P_1(\lambda)P_2(\lambda)$. *If P resolves completely into s parts, then $\lambda = 1$ is an s-tuple root of $P(\lambda) = 0$; and $s - 1$ of the components u_i are undetermined.*

(3) Consider the structures $P = \left(\begin{smallmatrix} P_1 & B \\ 0 & P_2 \end{smallmatrix}\right)$, or $P = \left(\begin{smallmatrix} P_1 & 0 \\ B & P_2 \end{smallmatrix}\right)$, where P_1 and P_2 are again square matrices, while B is not square, in general. [Often the standard structures, here and in example (2) can be identified only after renumbering the states.] If in the first type of structure, P_1 contains the states $1, 2, \ldots, r_1$, briefly: "The states of S_1", while P_2 contains the states of S_2: $r_1 + 1, \ldots, r$, then clearly, there is no transition from any state of S_1 to a state of S_2; a state of S_1 stays in S_1, while from S_2 to S_1 transition is possible. The states of S_2 are called *transient*, those of S_1 *permanent*. Give an example! Sketch the state diagram! *The stationary distribution for such a "reducible" matrix shows $u_i = 0$ for all transient states.*

(4) The simplest example of a *singular* or *periodic* P is the so-called *cyclic* matrix where $p_{i+1,i} = 1$, $i = 1, 2, \ldots, r - 1$; $p_{1r} = 1$; all other p_{ij} zero. The recurrence relations are here $p_i^{(m+1)} = p_{i-1}^{(m)}$, $i = 2, 3, \ldots, r$, $p_1^{(m+1)} = p_r^{(m)}$. They show that the $\mathbf{p}^{(m)}$ *oscillate with period r*. The reader may study in detail the cyclic matrix with $r = 3$, and compute P^m, $m = 1, 2, \ldots$; take $p_1^{(0)} : p_2^{(0)} : p_3^{(0)} = a : b : c$ and find the $p_i^{(m)}$.

It may easily be verified that *for the cyclic matrix of order r the characteristic roots are the r roots of unity.*

12.4. *More on classification.* A square matrix (p_{ij}), $i, j = 1, 2, \ldots, r$ is called *irreducible* if for any partition of the r numbers $1, 2, \ldots, r$ into two mutually exclusive and exhaustive parts, S_1 and $S - S_1 = S_2$, there is at least one positive p_{ij} such that i belongs to S_1 and j to S_2. The matrices of examples (1) and (4) are irreducible, those of (2) and (3) reducible. The characteristic property of an irreducible matrix is that *for any two subscripts i, k, an m exists such that $p_{ik}^{(m)} > 0$.*

For any irreducible P a unique stationary u_i exists and if P is in addition non-singular the $p_i^{(m)}$ converge toward u_i. A sufficient condition for this *positive regular* case is that *for some m, all $p_{ij}^{(m)}$ are positive.*

The reducible matrices are either *completely reducible* as in Ex. (2) of p. 215 or *reducible* as in (3) above. In order to avoid unessential complications we assume that the square matrices P_1, P_2, ... which formed the "parts" of P in (2) and (3) are *not further reducible and non-singular*. Case (2) (completely reducible) is then Hadamard-Fréchet's semi-regular case.

The reducible matrix of Ex. (3) has uniquely determined u_i's (and those u_i which correspond to the transient states are zero). Hence *the reducible case is regular.*

Foregoing complete proofs we formulate here some facts which should be fairly obvious by our examples and discussions.

In the regular case, $\lambda = 1$ *is a simple root of* $P(\lambda) = 0$ *and there is no other root of absolute value one. If P is regular and irreducible it is positive regular: all* $u_i > 0$. *If P is reducible, but not completely reducible, it is regular and* $u_i > 0$, *for the solutions which correspond to the permanent states, while* $u_i = 0$, *for the solutions which correspond to the transient states.*

If P resolves completely into n (irreducible and non-singular) *parts,* $\lambda = 1$ *is an n-tuple latent root and there is no other root of absolute value one. The* $p^{(n)}$ *converge toward a limit distribution which depends on the initial distribution and on the stationary distributions of the parts.*

The irreducible matrices are either positive regular or singular. Any singular P has roots of unity other than one.

The *irreducible singular matrix* may be shown (possibly after renumbering states) to have the structure of a cyclic matrix with the non-zero terms, $p_{21}, p_{32}, ..., p_{1s}$ "replaced" by stochastic matrices $P_1, P_2, ..., P_s$.

The r states 1, 2, ..., *r of a singular irreducible chain resolve into s groups* $G_1, G_2, ..., G_s$ *such that one-step transitions are possible from* G_ν to $G_{\nu+1}$, $\nu = 1, 2, ..., s - 1$ *and from* G_s to G_1; *no other transitions between groups are possible.*

The $p_i^{(m)}$ *and* $p_{ij}^{(m)}$ *of a cyclic matrix of order s are strictly periodic with period s; for the above-described general structure, the* $p_i^{(m)}$ *and* $p_{ij}^{(m)}$ *are asymptotically periodic.*

The reader will have no difficulties to illustrate all these statements by means of examples.

13. *Applications of Markov Chains*

13.1. *Independent trials. Random walk. Card shuffling. Automaton.*

(a) *Independent trials.* The transition matrix consists of r identical columns: for all j, $p_{ij} = p_i$, $i = 1, 2, ..., r$. If the initial distribution consists of the r numbers $q_1, q_2, ..., q_r$ with sum one we have $p_i^{(1)} = p_i q_1 + p_i q_2 + \cdots + p_i q_r = p_i(q_1 + \cdots + q_r) = p_i$, $i = 1, 2, ..., r$. Thus $p_i^{(m)} = p_i$, $i = 1, 2, ..., r$ for all m.

(b) *Random walk with absorbing barriers.* Consider the following one-dimensional random walk of a particle. Possible positions are the points with abscissas a, $a + 1$, $a + 2$, ..., $a + r - 1 = b$ of a line segment. At given (equidistant) times $t_0, t_1, t_2, \ldots$, the particle suffers shocks and changes position. At $x = a$ and $x = b$ there are *absorbing barriers* such that from a and b no transition occurs to any other state.

There are transition probabilities p to the right and q to the left from each of the $r - 2$ interior points. Accordingly, the non-vanishing transition probabilities are

$$p_{11} = 1, \qquad p_{rr} = 1, \qquad p_{i,i+1} = q, \qquad i = 1, 2, \ldots, r - 2$$
$$p_{i,i-1} = p, \qquad i = 3, 4, \ldots, r \tag{94}$$

The reader may sketch the state diagram. The transition matrix has $\lambda = 1$ as a double root. The matrix resolves, but not completely, into two parts. One consists of the states 1 and r, which form a completely reducible matrix with $p_{11}^{(n)} = 1$, $p_{rr}^{(n)} = 1$ for all n. The other consists of the $r - 2$ inner states, which are transient, since from each such state there is a non-zero chance of reaching an absorbing barrier eventually. The stationary distribution is of the form: $u_1 > 0$, $u_2 = u_3 = \cdots = u_{r-1} = 0$, $u_r > 0$, $u_1 + u_r = 1$; the values of u_1 and of u_r depend on the initial distribution.

(c) *Card shuffling.* A deck of q cards can be arranged in $r = q!$ ways. Each of the r permutations of the q numbers is considered a state.[1] By a single operation of shuffling one state is transformed into another. Let us assume that each shuffling operation S_j has a certain probability p_j where $p_j > 0$, $\Sigma_{j=1}^{r} p_j = 1$. Suppose an initial state given. By the operation S_i it is transformed into a state (a permutation) which we denote likewise by S_i. Denote then by S_jS_i the operation of first applying the shuffle S_i and then S_j and by S_i^{-1} the operation inverse to S_i. Then $S_iS_k^{-1}$ is the operation which, applied to the state S_k transforms it into the state S_i; let p_{ik} be the probability of this operation. Each p_{ik} is equal to some p_j since $S_iS_k^{-1}$ is equivalent to some S_j. Therefore, the matrix of the p_{ik} which corresponds to $S_iS_k^{-1}$, $i, k = 1, 2, \ldots, r$ contains in each line and in each column all S_j, though in a different order. It follows that

$$\sum_{k=1}^{r} p_{ik} = \sum_{j=1}^{r} p_j = 1, \qquad \sum_{k=1}^{r} p_{ki} = 1, \qquad i = 1, 2, \ldots, r. \tag{95}$$

This property of P is sometimes called *doubly stochastic.*

[1] H. Poincaré, *Calcul des Probabilités,* 2ème éd., Chapter XVI. Paris, 1912; J. Hadamard, "Sur le battage des cartes." *C. R. Acad. Sci. (Paris)* **185** (1927), pp. 5–9.

If all p_{ik} are positive, the problem is positive regular; then $p_{ik}^{(m)} \to u_i$, and it follows from (95) that $u_1 = u_2 = \cdots = u_r = 1/r = 1/q!$:

$$\lim_{m\to\infty} p_{ik}^{(m)} = \frac{1}{q!}\,. \tag{96}$$

Hence, *under these assumptions any given state will appear after infinitely many shufflings with the same probability* $1/q!$ *which is independent of the initial state.*[2] This corresponds to the idea of "thorough mixing."

(d) *Finite automatons.* Problems related to finite automata correspond to a particular type of Markov chain.

There is a (finite) number of states A_1, A_2, ..., A_r. The automaton receives a sequence of signals 1, 2, ..., s which dirèct it to change from one state to another state. At each of the r states any of the s signals appear with probabilities a_i, $i = 1, 2, ..., s$, $\sum_{i=1}^{s} a_i = 1$.

Consider an example with $r = 4$, $s = 6$. The state diagram is shown in Fig. 13. There is at state A_1 a probability a_1 for the system to remain in

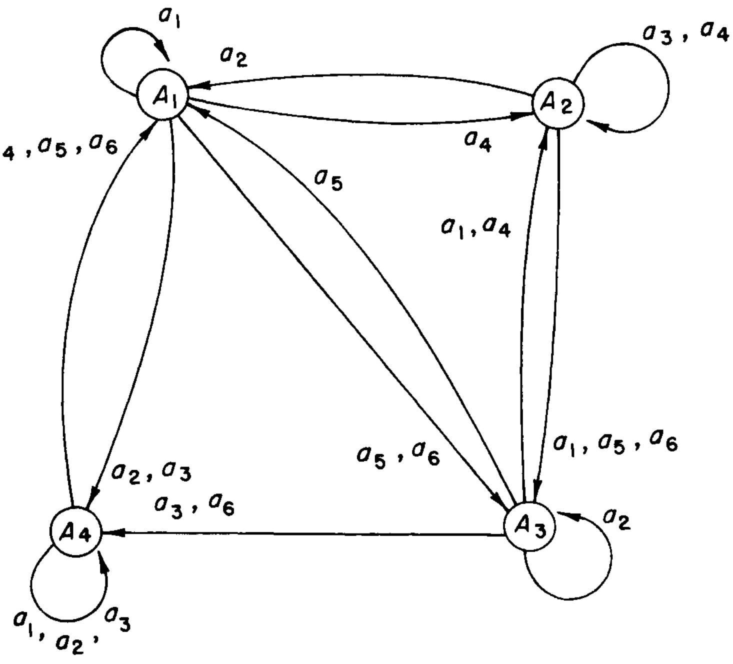

FIG. 13. A finite automaton.

[2] One can devise methods of shuffling where the corresponding P is, for example, completely reducible, or cyclic. All depends on the p_{ik}.

state A_1, a probability a_4 to change from A_1 to A_2, a probability $a_5 + a_6$ for a change from A_1 to A_3, and probability $a_2 + a_3$ for the change from A_1 to A_4, etc.

The corresponding state matrix A is

$$A = \begin{pmatrix} a_1 & a_2 & a_5 & a_4 + a_5 + a_6 \\ a_4 & a_3 + a_4 & a_1 + a_4 & 0 \\ a_5 + a_6 & a_1 + a_5 + a_6 & a_2 & 0 \\ a_2 + a_3 & 0 & a_3 + a_6 & a_1 + a_2 + a_3 \end{pmatrix}.$$

It is characteristic of this type of matrix that the $r^2 - r$ independent probabilities p_{ij} of a general stochastic r-matrix are all expressed here in terms of $s - 1$ independent probabilities a_i, $i = 1, 2, ..., s$.

An automaton can be described in terms of one single "urn" which contains the labels $1, 2, ..., s$ in the proportion $a_1 : a_2 : \cdots : a_s$. The game starts at a given state, A_2, say $[p_2^{(0)} = 1$, all other $p_i^{(0)} = 0]$; we draw a lot which reads 3, say; therefore (in our example) the system remains in A_2. After replacement we draw again and obtain 2; therefore the system moves to A_1, etc.

On the other hand, *any finite Markov chain with rational probabilities can be interpreted, and in various ways, as a finite automaton,* where the number of possible signals is equal to or smaller than the lowest common multiple of the denominators of the p_{ij}. Consider for example in (97) the stochastic matrix P with $r = 3$:

$$P = \begin{pmatrix} \frac{1}{2} & \frac{5}{12} & 0 \\ \frac{1}{4} & \frac{7}{12} & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{1}{2} \end{pmatrix}, \tag{97}$$

$$A = \begin{pmatrix} a_1 + \cdots + a_6 & a_1 + \cdots + a_5 & 0 \\ a_7 + a_8 + a_9 & a_6 + \cdots + a_{12} & a_3 + \cdots + a_8 \\ a_{10} + a_{11} + a_{12} & 0 & a_1 + a_2 + a_9 + \cdots + a_{12} \end{pmatrix}. \tag{97'}$$

We may interpret P in terms of 12 signals all with equal probability $a_i = \frac{1}{12}$, $i = 1, ..., 12$ and arrive at the state matrix A of (97′). We may, however, equally well associate to P a state matrix B with five signals of probabilities $b_1 = \frac{1}{6}$, $b_2 = \frac{1}{3}$, $b_3 = \frac{1}{4}$, $b_4 = \frac{1}{12}$, $b_5 = \frac{1}{6}$ and obtain the state matrix

$$B = \begin{pmatrix} b_1 + b_2 & b_3 + b_5 & 0 \\ b_3 & b_1 + b_2 + b_4 & b_1 + b_3 + b_4 \\ b_4 + b_5 & 0 & b_2 + b_5 \end{pmatrix}. \tag{97''}$$

The various concepts and results valid for stochastic matrices apply to the finite automata. We may, for example, consider a reducible P with $r = 6$, A_1, A_2 transient, A_3, A_4, A_5, A_6 permanent, and the stable distribution $u_1 = u_2 = 0$, $u_i > 0$, $i = 3, \ldots, 6$. A typical question would be to ask how large m ought to be for $p_1^{(m)} + p_2^{(m)}$ to become sufficiently small that one may be justified in neglecting these transient states.

13.2. *Two-state Markov chain. Law of large numbers.* This case is instructive since explicit formulas can be given. Modifying the notation of Section 12 we put

$$p_1^{(0)} = p_1, \quad p_2^{(0)} = 1-p_1, \quad p_{11} = \alpha, \quad p_{12} = \beta, \quad p_{21} = 1-\alpha, \quad p_{22} = 1-\beta.$$

We assume $| \alpha - \beta | \neq 1$, i.e., neither $\alpha = 1, \beta = 0$ nor $\alpha = 0, \beta = 1$, these cases being exceptions of little interest. We also introduce the notation

$$p_1^{(n)} = p_n; \qquad p_{11}^{(n)} = p^{(n)}, \qquad p_{12}^{(n)} = q^{(n)}.$$

The roots of P are $\lambda_1 = 1$, $\lambda_2 = \alpha - \beta = \delta$.[3] Under our hypothesis, the transition matrix is positive regular and the p_n, $1 - p_n$ converge toward a positive stationary distribution u, $1 - u$ for which we write here p and $1 - p$. We have

$$u = p = \lim_{n\to\infty} p_n = \frac{\beta}{1+\beta-\alpha} = \frac{\beta}{1-\delta},$$

$$1 - u = 1 - p = \frac{1-\alpha}{1+\beta-\alpha} = \frac{1-\alpha}{1-\delta}. \tag{98}$$

Formulas (87) of Section 12 give for p_n the recurrence relation

$$p_n = \alpha p_{n-1} + \beta(1 - p_{n-1}) = \delta p_{n-1} + \beta. \tag{99}$$

The solution of this difference equation of first order with initial value p_1 is

$$p_n = \frac{\beta}{1-\delta} + \left(p_1 - \frac{\beta}{1-\delta}\right)\delta^{n-1} = p + (p_1 - p)\delta^{n-1}, \tag{99'}$$

and (excluding again the cases $\delta = 1$, $\delta = -1$)

$$\bar{p}_n = \frac{1}{n}(p_1 + p_2 + \cdots + p_n) = p + \frac{p_1 - p}{n}\frac{1-\delta^n}{1-\delta}, \tag{99''}$$

which is seen to converge toward p as $n \to \infty$.

[3] Note that δ is also the value of the determinant $\alpha(1 - \beta) - \beta(1 - \alpha)$ of the transition matrix, which under our assumption is less than one.

The recurrence equations for the $p^{(n)}$, $q^{(n)}$ are similar to (99), viz.,

$$p^{(n)} = \alpha p^{(n-1)} + \beta[1 - p^{(n-1)}]; \quad q^{(n)} = \alpha q^{(n-1)} + \beta[1 - q^{(n-1)}] \tag{100}$$

with the initial conditions $p^{(1)} = \alpha$ and $q^{(1)} = \beta$. The solutions are

$$p^{(n)} = p + (1 - p)\delta^n, \qquad q^{(n)} = p - p\delta^n. \tag{100'}$$

If, in the usual way, we associate the number $x_\nu = 1$, with "occurrence" of the "event" in the νth trial, and $x_\nu = 0$ with non-occurrence, then $E[x_\nu] = E[x_\nu{}^2] = p_\nu$ and

$$E[x_1 + x_2 + \cdots + x_n] = E[x] = p_1 + p_2 + \cdots + p_n\,,$$

where x is the number of successes in n trials. Hence using (99″)

$$E\left[\frac{x}{n}\right] = \frac{p_1 + p_2 + \cdots + p_n}{n} = \bar{p}_n = p + O\left(\frac{1}{n}\right). \tag{101}$$

Next, we compute the variance of $x_1 + \cdots + x_n$. Clearly,

$$E[x_\nu{}^2] = p_\nu\,, \qquad E[x_\nu x_\mu] = p_\nu p^{(\mu-\nu)}, \qquad \mu > \nu$$

$$E[(x_\nu - p_\nu)^2] = p_\nu - p_\nu{}^2 = pq + (q - p)(p_1 - p)\delta^{\nu-1} - (p_1 - p)^2\delta^{2\nu-2}$$

$$E[(x_\nu - p_\nu)(x_\mu - p_\mu)] = E[x_\nu x_\mu] = p_\nu p_\mu = p_\nu p^{(\mu-\nu)} - p_\nu p_\mu$$

$$= pq\delta^{\mu-\nu} + (p_1 - p)(q - p)\delta^{\mu-1} - (p_1 - p)^2\delta^{\nu+\mu-2}.$$

Now consider

$$s_n{}^2 = E[(x_1 + \cdots + x_n - p_1 - \cdots - p_n)^2] = \sum_{\nu=1}^{n} E[(x_\nu - p_\nu)^2]$$

$$+ 2\sum_{\mu>\nu} E[(x_\nu - p_\nu)(x_\mu - p_\mu)] = A + 2B,$$

where

$$A = \sum_{\nu=1}^{n} E[(x_\nu - p_\nu)^2] = npq + (q - p)(p_1 - p)\frac{1 - \delta^n}{1 - \delta} - (p_1 - p)^2\frac{1 - \delta^{2n}}{1 - \delta^2}\,.$$

Here the last two terms remain bounded as $n \to \infty$. To find B we let first ν run over $1, 2, \ldots, n - 1$, then $\mu = \nu + 1, \nu + 2, \ldots, n$. We take first the μ-summation and find for the first term

$$pq(\delta + \delta^2 + \cdots + \delta^{n-\nu}) = pq\frac{\delta - \delta^{n-\nu+1}}{1 - \delta}\,.$$

Proceeding in the same way for the other terms we have

$$pq\frac{\delta - \delta^{n-\nu+1}}{1-\delta} + (p_1 - p)(q - p)\frac{\delta^{\nu} - \delta^{n}}{1-\delta} - (p_1 - p)^2\delta^{\nu}\frac{\delta^{\nu-1} - \delta^{n-1}}{1-\delta}$$

Now we take $\nu = 1, 2, 3, \ldots, n-1$:

$$B = \sum_{\mu > \nu} E[(x_\nu - p_\nu)(x_\mu - p_\mu)] = npq\frac{\delta}{1-\delta} - \frac{pq}{1-\delta}\delta^2\frac{1-\delta^{n-1}}{1-\delta} + \cdots .$$

It is seen that all terms of B except the first one remain bounded as $n \to \infty$ and we obtain asymptotically

$$s_n{}^2 = A + 2B \sim npq\frac{1+\delta}{1-\delta}. \qquad (102)$$

From this and (101) we have the result: *In the case of a simple chain with* $r = 2$, the distribution of $z = (x_1 + x_2 + \cdots + x_n)/n$, *the average number of events, has asymptotically the mean value* p *and the variance* $\frac{p(1-p)}{n}\;\frac{1+\delta}{1-\delta}$ where $p = \lim_{n\to\infty} p_n$. In Chapter VI we shall consider an application of these results to the theory of the Galton board.

From (102) we see that

$$\lim_{n\to\infty}\frac{s_n{}^2}{n^2} = 0.$$

As in Section 4, this implies a law of large numbers, here for Markov chains: $z - \bar{p}_n$, and likewise $z - p$, *converge in probability to zero as* $n \to \infty$. Or, equivalently,

$$\Pr\{|\, z - p\,| < \epsilon\} \geqslant 1 - \frac{s_n{}^2}{n^2\epsilon^2}. \qquad (103)$$

It is not difficult to obtain the analog for general r.