Chapter 2

Classical Probability Theories

In principle, those who are not interested in mathematical foundations of probability theory might jump directly to Part II. One should just know that, besides the Kolmogorov definition of classical probability as a probability measure, another classical probability theory was developed by von Mises: probability was defined as the limit of relative frequencies.

We are well aware of mathematical difficulties of the von Mises theory, cf., e.g., [305, 315, 300, 322, 234, 139]. These difficulties are consequences of von Mises' definition of randomness through so-called *place selections.* There are few ways to escape these difficulties. One way is to follow von Mises and choose a class of place selections depending on a series of experiments under consideration, compare also with Wald [315]. Another way for development was proposed by Kolmogorov. This is *complexity theory* for random sequences: [223, 224, 50, 51, 288–290, 322, 304, 234, 282]. We also mention the theory of recursive statistical tests (initiated by Kolmogorov and developed by Martin-Löf [244, 245]).

Besides these two mathematically advanced ways, it is possible to choose a very pragmatic way of modification of von Mises's theory to obtain a rigorous mathematical formalism. At the moment the problem of randomness of sequences of experimental data is not of great interest in quantum physics. The quantum experimental research is not yet so much devoted to randomness of statistical data. It is devoted merely to the study of relative frequencies (which stabilize to probabilities in long runs of experiments). Therefore we can forget (at least for a while) about von Mises' attempt to provide a mathematical formalization of the notion of randomness.

We shall proceed by considering only statistical stabilization of relative frequencies—*existence of the limit of a sequence of relative frequencies.* Thus we shall enjoy all advantages of the von Mises frequency approach to probability and at the same time we shall escape all difficulties related to a rigorous definition of randomness.

2.1 Kolmogorov Measure-Theoretic Model

The axiomatics of modern probability theory were proposed by Andrei Nikolaevich Kolmogorov [222] in 1933.

2.1.1 Formalism

We recall some notions of measure theory. Let Ω be a set. A system F of subsets of a set Ω is called an *algebra* if the sets $\emptyset$, Ω belong to F and the union, intersection and difference of two sets of F also belong to F. In particular, for any $A \in F$, *complement* $\bar{A} = \Omega \setminus A$ of A belongs to F.

Denote by F_Ω the family of all subsets of Ω. This is the simplest example of an algebra.

Let F be an algebra. A map $\mu : F \to \mathbf{R}_+$ is said to be a *measure* if $\mu(A \cup B) = \mu(A) + \mu(B),$ for $A, B \in F, A \cap B = \emptyset$. A measure μ is called σ*-additive* if, for every sequence $\{A_n\}_{n=1}^\infty$ of sets $A_n \in F$ such that their union $A = \bigcup_{n=1}^\infty A_n$ also belongs to F, we have: $\mu(A) = \sum_{n=1}^\infty \mu(A_n)$. An algebra, say $\mathscr{F}$, is said to be a σ-*algebra* if, for every sequence $\{A_n\}_{n=1}^\infty$ of sets $A_n \in \mathscr{F}$, their union $A = \bigcup_{n=1}^\infty A_n$ belongs to $\mathscr{F}$.

Some books on probability theory use the terminology *field* and σ*-field* of sets, instead of algebra and σ-algebra.

Let Ω_1, Ω_2 be arbitrary sets and let G_1, G_2 be some systems of subsets of Ω_1 and Ω_2, respectively. A map $\xi : \Omega_1 \to \Omega_2$ is called *measurable* or more precisely $((\Omega_1, G_1), (\Omega_2, G_2))$-measurable if, for any set $A \in G_2$, the set $\xi^{-1}(A) \in G_1$. Here $\xi^{-1}(A) = \{\omega \in \Omega_1 : \xi(\omega) \in A\}$. We shall use the notation $\xi : (\Omega_1, G_1) \to (\Omega_2, G_2)$ to indicate dependence on G_1, G_2. Typically we shall consider measurability of maps in the case such that the systems of sets G_j, $j = 1, 2$, are algebras or σ-algebras.

Let A be a set. The *characteristic function* I_A of the set A is defined as $I_A(x) = 1$, $x \in A$, and $I_A(x) = 0$, $x \in \bar{A}$.

Let $A = \{a_1, \dots, a_n\}$ be a finite set. We shall denote the number of elements n of A by the symbol $|A|$.

By the Kolmogorov axiomatics [222], see also [286], a *probability space* is a triple

$$\mathscr{P} = (\Omega, \mathscr{F}, \mathbf{P}),$$

where Ω is an arbitrary set (points ω of Ω are said to be *elementary events*), $\mathscr{F}$ is an arbitrary σ-algebra of subsets of Ω (elements of $\mathscr{F}$ are said to be *events*), $\mathbf{P}$ is a σ-additive measure on $\mathscr{F}$ which yields values in the segment [0, 1] of the real line and normalized by the condition $\mathbf{P}(\Omega) = 1$ (it is said to be *probability*).

Random variables on the Kolmogorov space $\mathscr{P}$ are by definition measurable functions $\xi : \Omega \to \mathbf{R}$, or in our notation $\xi : (\Omega, \mathscr{F}) \to (\mathbf{R}, \mathscr{B})$), where $\mathscr{B}$ is the Borel σ-algebra on the real line.[1] We shall use the symbol $RV(\mathscr{P})$ to denote the space of random variables for the probability space $\mathscr{P}$. The probability distribution of $\xi \in RV(\mathscr{P})$ is defined by $\mathbf{P}_\xi(B) = \mathbf{P}(\xi^{-1}(B))$ for $B \in \mathscr{B}$. This is a σ-additive

[1] Thus $\xi^{-1}(B) \in \mathscr{F}$ for every $B \in \mathscr{B}$.

measure on the Borel σ-algebra. The *average* (mathematical expectation) of a random variable ξ is defined by

$$E\xi = \int_{\Omega} \xi(\omega) d\mathbf{P}(\omega) \tag{2.1}$$

In particular, if $\xi = a_1, \ldots, a_n, \ldots$ is a discrete random variable its average is given by

$$E\xi = \sum_n a_n \mathbf{P}(\xi = a_n). \tag{2.2}$$

Conditional probability will play an essential role in further quantum considerations. In Kolmogorov's probability model *conditional probability* is defined by well-known *Bayes' formula*. In many textbooks this formula is called Bayes' theorem. However, in the Kolmogorov model it is neither a theorem nor an axiom, but a *definition.* Conditional probability is introduced in the Kolmogorov model through the following definition:

$$\mathbf{P}(B|A) = \mathbf{P}(B \cap A)/\mathbf{P}(A), \qquad \mathbf{P}(A) > 0. \tag{2.3}$$

By Kolmogorov's interpretation this is the *probability that an event A occurs under the condition of an event B having occurred.*

The *Kolmogorov probability model* is given by the Kolmogorov probability space endowed with conditional probability via the Bayes formula.

We remark that $\mathbf{P}_A(B) \equiv \mathbf{P}(B|A)$ is again a probability measure on $\mathscr{F}$. The conditional expectation of a random variable ξ is defined by

$$E(\xi|A) = \int_{\Omega} \xi(\omega) d\mathbf{P}_A(\omega). \tag{2.4}$$

In the Kolmogorov model two events A and B are said to be *independent* if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)P(B) \tag{2.5}$$

or

$$\mathbf{P}(B|A) = P(B), \qquad \mathbf{P}(A) > 0. \tag{2.6}$$

In our further considerations an important role will be played by the *formula of total probability* - a theorem of Kolmogorov's model. Let us consider a countable family of sets $A_k \in \mathscr{F}, \mathbf{P}(A_k) > 0, k = 1, \ldots,$ such that

$$\bigcup_{k=1}^{\infty} A_k = \Omega, \quad \text{and} \quad A_k \cap A_l = \emptyset, \quad k \neq l.$$

Such a family is called a *measurable partition* of the space Ω or a *complete group of disjoint events*.

Theorem 2.1. *Let* $\{A_k\}$ *be a complete group of disjoint events. Then, for every set* $B \in \mathscr{F}$, *the following formula of total probability holds*:

$$\mathbf{P}(B) = \sum_{k=1}^{\infty} \mathbf{P}(A_k)\mathbf{P}(B|A_k). \tag{2.7}$$

Proof. We have

$$\mathbf{P}(B) = \mathbf{P}\left(B \cap \bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(B \cap A_k) = \sum_{k=1}^{\infty} \mathbf{P}(A_k)\frac{\mathbf{P}(B \cap A_k)}{\mathbf{P}(A_k)}.$$

Especially interesting for us is the case such that a complete group of disjoint events is induced by a discrete random variable a taking values $\{\alpha_k\}$. Here

$$A_k = \{\omega \in \Omega : a(\omega) = \alpha_k\}. \tag{2.8}$$

Let b be another random variable. It takes values $\{\beta_j\}$. Then we have:

$$\mathbf{P}(b = \beta) = \sum_{\alpha} \mathbf{P}(a = \alpha)\mathbf{P}(b = \beta|a = \alpha). \tag{2.9}$$

2.1.2 Discussion

The Kolmogorov axiomatics [222] played a crucial role in creation of the rigorous mathematical formalism of probability theory, see, e.g., [286, 285]. The main problem of this approach to probability is that by starting with an abstract Kolmogorov space $\mathscr{P} = (\Omega, \mathscr{F}, \mathbf{P})$ we completely lose information on the intrinsic structure of a physical statistical experiment. The main idea of Kolmogorov was that it is possible to proceed in a very abstract way without having to pay attention to structures of concrete ensembles:

"To outline the context of theory, it suffices to single out from probability theory those elements that bring out its intrinsic logical structure, but have nothing to do with the specific meaning of theory."

In quantum physics we cannot use one fixed Kolmogorov probability space $\mathscr{P} = (\Omega, \mathscr{F}, \mathbf{P})$—an *absolute Kolmogorov space*—to describe different experiments.[2] This impossibility was considered as a contradiction between the classical and quantum probabilistic descriptions.

Of course, it would be better if from the very beginning A. N. Kolmogorov defined a probability space as *a collection of conventional Kolmogorov probability spaces corresponding to a family of contexts (complexes of experimental physical conditions)* $\mathscr{C} = \{C\}$:

$$\{\mathscr{P}_C : C \in \mathscr{C}\}, \quad \text{where } \mathscr{P}_C = (\Omega_C, \mathscr{F}_C, \mathbf{P}_C). \tag{2.10}$$

Unfortunately, A. N. Kolmogorov did not do this, cf. S. Gudder [106–108]: theory of probabilistic manifolds. In our approach a structure which is similar to (2.10) will appear as a special case of the general contextual probability space, see Part II, Chap. 6 (Definition 6.8). It will be called the contextual multi-Kolmogorovian probability space. We remark that such a probabilistic construction does not contradict the original ideas of Kolmogorov [222] who noticed that each complex of experimental conditions generates its own (Kolmogorov) probability space. Thus operation with data collected for a few different complexes of, e.g., physical conditions

[2] Cf. with attempts to use absolute Newton space in classical physics.

should naturally induce families of (Kolmogorov) probability spaces—contextual multi-Kolmogorovian spaces. As was pointed out, although Kolmogorov emphasized this ideologically, he did not proceed in this direction mathematically.

2.2 Von Mises Frequency Model

Let us recall the main notions of a frequency theory of probability [309–311] of Richard von Mises (1919).[3]

2.2.1 Collective (Random Sequence)

Von Mises' probability theory is based on the notion of a *collective*. Consider a random experiment. Let a be some observable representing results of this random experiment. The set of all possible results of this experiment is $L = \{s_1, \ldots, s_m\}$—the label set or the set of attributes. It will be finite in this book.

Consider N observations of a and write a result x_j after each trial. We obtain the finite sample: $x = (x_1, \ldots, x_N),\ x_j \in L$. A *collective* is an infinite idealization of this finite sample:

$$x = (x_1, \ldots, x_N, \ldots), \quad x_j \in L, \tag{2.11}$$

for which two von Mises' principles are valid. Let us compute frequencies

$$\nu_N(s; x) = \frac{n_N(s; x)}{N}, \quad s \in L,$$

where $n_N(s; x)$ is the number of realizations of the attribute s in the first N trials.

[3] In fact, already in 1866 John Venn tried to define a probability explicitly in terms of relative frequencies.

Principle S (Statistical stabilization of relative frequencies). *For every label* $s \in L$, *the frequency* $\nu_N(s; x)$ *approaches a limit as* N *approaches infinity.*

In the frequency theory of probability the limit $\mathbf{P}(s) = \lim \nu_N(s; x)$ is the probability of the label s. Sometimes this probability will be denoted by $\mathbf{P}_x(s)$ (to show a dependence on the collective x).

Principle R (Randomness). *The limits of relative frequencies have to be stable with respect to a place selection (a choice of a subsequence) in* (2.11).

Heuristically it is evident that we cannot consider, for example, the sequence $z = (0, 1, 0, 1, \ldots, 0, 1, \ldots)$ as a random object (generated by a statistical experiment). Principle **S** holds for z and $\mathbf{P}(0) = \mathbf{P}(1) = 1/2$. But, the sequence z does not satisfy Principle **R**. If we choose only even places, then we obtain the zero sequence $z_0 = (0, 0, \ldots)$, where $\mathbf{P}(0) = 1, \mathbf{P}(1) = 0$.

The *average* of observable a is defined as the average with respect to the probability distribution $\mathbf{P}_x$:

$$E_x a = \sum_{s \in L} s \mathbf{P}_x(s). \tag{2.12}$$

Here x is a collective representing observations of a.

Finally, we recall the original von Mises thoughts about the notion of collective:

"We will say that a collective is a mass phenomenon or a repetitive event, or simply a long sequence of observations for which there are sufficient reasons to believe that the relative frequency of the observed attribute would tend to a fixed limit if the observations were infinitely continued. This limit will be called the probability of the attribute considered within the given collective," R. von Mises [311].

2.2.2 Difficulties with Definition of Randomness

However, this very natural notion of randomness was the hidden bomb in the foundations of von Mises' theory. The main problem was to define a class of place selections which would induce a fruitful theory. A very natural restriction is that a place selection in (2.11) cannot be based on the use of attributes of elements. For example, we cannot consider a subsequence of (2.11) constructed by choosing elements with the fixed label $s \in L$. Von Mises proposed the following definition of a place selection:

(PS) "a subsequence has been derived by a place selection if the decision to retain or reject the nth element of the original sequence depends on the number n and on label values $x_1, \dots, x_{n-1}$ of the $(n-1)$ presiding elements, and not on the label value of the nth element or any following element", see [87, p. 9].

Thus a place selection can be defined by a set of functions

$$f_1, \quad f_2(x_1), \quad f_3(x_1, x_2), \quad f_4(x_1, x_2, x_3), \quad \dots,$$

each function yielding the values 0 (rejecting the nth element) or 1 (retaining the nth element). There are some examples of place selections: (1) choose those x_n for which n is prime; (2) choose those x_n which follow the word 01; (3) toss a (different) coin; choose x_n if the nth toss yields heads.

The first two selection procedures may be called *lawlike*, the third random. It is more or less obvious that all of these procedures are place selections: the value of x_n is not used in determining whether to choose x_n.

The principle of randomness ensures that no strategy using a place selection rule can select a subsequence that allows different odds for gambling than a sequence that is selected by flipping a fair coin. This principle can be called the *law of excluded gambling strategy.*

The definition (PS) induced some mathematical problems. If a class of place selections is too extended, then the notion of the collective is too restricted (in fact, there are no sequences where probabilities are invariant with respect to all place selections). This was the main critical argument against von Mises' theory.

However, von Mises himself was totally satisfied by the following operational solution of this problem. He proposed [311] to fix for any collective a class of place selections which depends on the physical problem described by this collective. Thus he moved the problem outside the mathematical framework. For any concrete experiment, one should find a special class of place selections which would be appropriative for this experiment.

2.2.3 S-sequences

As probability is defined on the basis of the principle of the statistical stabilization of relative frequencies, it is possible to develop a quite fruitful probabilistic calculus *based only on this principle.* Instead of the notion of a collective, we can consider a more general notion.

Definition 2.1. *A sequence x, see (2.11), which satisfies the principle of the statistical stabilization of relative frequencies is said to be an S-sequence.*

Thus the limits of relative frequencies in an S-sequence x need not be invariant with respect to some class of place selections.

It seems that the machinery of randomness has no applications in quantum physics. Experimenters are only interested in the statistical stabilization of relative frequencies.

2.2.4 Operations for Collectives

On many occasions R. von Mises emphasized that frequency probability theory is not a calculus of probabilities, but is the *calculus of collectives* which generates the corresponding calculus of probabilities. We briefly discuss some of the basic operations for collectives (see [311] for the details) and S-sequences (see [139] for details).

(a) **Operation of mixing of labels in an S-sequence (or collective) and additivity of probabilities.** Let x be an S-sequence (in particular, it can be a collective) with the (finite) label space $L_x = \{s_1, \ldots, s_m\}$:

$$x = (x_1, \ldots, x_N, \ldots), \quad x_j \in L_x, \tag{2.13}$$

and let $E = \{s_{i_1}, \ldots, s_{i_d}\}$ be a subset of the set of labels L_x. The sequence (2.13) of x is transformed into a new sequence y_E by the following rule:

If x_j belongs to the set E, then we write 1; if x_j does not belong to the set E then we write 0.

Thus the label set of the sequence y_E constructed on the basis of this rule is $L_{y_E} = \{0, 1\}$. R. von Mises called this operation on sequences, $x \to y_E$, the operation of mixing of labels. We take a subset E of the label set L_x and we "mix" elements of E into a new label, $y = 1$; elements of the complement of E are mixed into the label $y = 0$.

Proposition 2.1. *If a sequence of labels x satisfies the principle of statistical stabilization (so it is an S-sequence), then, for any subset E of the label set L_x, the sequence y_E also satisfies the principle of statistical stabilization (so it is also an S-sequence).*

Proof. For example, for the label 1 we have:

$$\mathbf{P}_{y_E}(1) = \lim_{N\to\infty} \nu_N(E; x) = \lim_{N\to\infty} \sum_{k=1}^{d} \nu_N(s_{i_k}; x) = \sum_{k=1}^{d} \mathbf{P}_x(s_{i_k}), \tag{2.14}$$

where $\nu_N(E; x) \equiv \nu_N(1; y_E) = n_N(1; y_E)/N$ is the relative frequency of 1 in y_E. To obtain (2.14) we have only used the fact that the addition is a continuous operation on the field of real numbers $\mathbf{R}$.

We can also show that if a sequence x satisfies the principle of randomness (so it is a collective), then the sequence y_E also satisfies the principle of randomness (so it is also a collective), see [311, 139].

By this operation any S-sequence (in particular, any collective) x generates a probability distribution on the algebra F_{L_x} of all subsets of the label set L_x. By definition we have:

$$\mathbf{P}_x(E) = \mathbf{P}_{y_E}(1) = \sum_{s \in E} \mathbf{P}_x(s).$$

Now we find the properties of this probability. We start with a simple, but extremely important result.

Theorem 2.2. *For any S-sequence x, the frequency probability $\mathbf{P}_x$ yields values in the segment* $[0, 1]$.

Proof. As $\mathbf{P}_x(E) = \lim_{N\to\infty} \nu_N(E; x)$ and $0 \le \nu_N(E) \le 1$, then (by an elementary theorem of real analysis)

$$0 \le \mathbf{P}_x(E) \le 1.$$

We emphasize that in frequency probability theory this is a theorem and not an axiom (as it is in the Kolmogorov measure-theoretic model). Another very simple, but extremely important result is also an axiom in the Kolmogorov model, but a theorem in the von Mises model.

Theorem 2.3. *For any S-sequence x, the frequency probability $\mathbf{P}_x$ is normalized by* 1.

Proof. As the S-sequence y_{L_x} corresponding to the whole label set L_x does not contain zeros, we obtain that for any N the relative frequency $\nu_N(L_x; x) \equiv \nu_N(1; y_{L_x}) \equiv 1$ and, consequently,

$$\mathbf{P}_x(L_x) = \sum_s \mathbf{P}_x(s) = 1. \tag{2.15}$$

Finally by (2.14) we find that the set function

$$\mathbf{P}_x : F_{L_x} \to [0, 1]$$

is additive. Thus we have obtained

Theorem 2.4. *The frequency probability* $\mathbf{P}_x$ *is additive.*

Thus $\mathbf{P}_x$ is a normalized measure on the set-algebra F_{L_x} which yields values in [0, 1].

(b) **Operation of partition of an S-sequence (or a collective) and conditional probabilities.** Let x be an S-sequence (or even a collective). We take a subset, say O, of the label set L_x such that $\mathbf{P}_x(O) \neq 0$. Thus $\mathbf{P}_x(s) = \lim_{N\to\infty} \nu_n(s; x) > 0$ for at least one label $s \in L_x$. We now derive a new sequence $z(O)$ by the following rule:

There are retained only those elements of x which belong to subset O and all other elements are discarded.

The label set of the sequence $z(O)$ coincides with the set O. This operation is obviously not a place selection, since the decision to retain or reject an element of x depends on the label of just this element.[4]

Proposition 2.2. *For any S-sequence x and any subset O of the label set L_x such that* $\mathbf{P}_x(O) \neq 0$, *the sequence $z(O)$ is again an S-sequence.*

Proof. Suppose that $s \in O$ and let y_O be the S-sequence generated by x with the aid of the mixing operation. Then

$$\mathbf{P}_{z(O)}(s) = \lim_{N\to\infty} \nu_N(s; z(O)) = \lim_{k\to\infty} \nu_{N_k}(s; z(O)),$$

where $N_k \to \infty$ is an arbitrary sequence. As $\mathbf{P}(O) \neq 0$, then

$$M_k = n_k(1; y_O) \to \infty.$$

[4] It is important to remark that this operation is not based on a new measurement. We just operate with data which was collected via a measurement represented by an S-sequence (collective) x; compare with considerations that will be presented in Part II, Chap. 1.

This is the number of labels belonging to O among the first k elements of x. Thus we have:

$$\mathbf{P}_{z(O)}(s) = \lim_{k\to\infty} \nu_{M_k}(s; z(O)) = \lim_{k\to\infty} n_{M_k}(s; z(O))/M_k$$
$$= \lim_{k\to\infty} [n_{M_k}(s; z(O))/k] : [M_k/k] = \mathbf{P}_x(s)/\mathbf{P}_x(O). \tag{2.16}$$

We have used the property that $n_{M_k}(s; z(O))$—the number of occurrences of the label s among first M_k elements of $z(O)$—is equal to $n_k(s; x)$—the number of occurrences of the label s among first k elements of x.

It is also possible to show that if x is a collective, then the sequence $z(O)$ is again a collective, see [309, 139].

Definition 2.2. *The probability $\mathbf{P}_{z(O)}(s)$ is called the conditional probability of the label $s \in L_x$ under the condition that it belongs to the subset O of the label set L_x.*

This probability is denoted by $\mathbf{P}_x(s|O)$. For any $B \subset O$, we define conditional probability by

$$\mathbf{P}_x(B|O) = \mathbf{P}_{z(O)}(B) = \sum_{s\in B} \mathbf{P}_x(s|O).$$

Sometimes we shall use the symbol $\mathbf{P}(B|O)$. However, if we forget that, in fact, the probability $\mathbf{P}$ depends on a collective x, then the symbol $\mathbf{P}(B|O)$ might induce misunderstanding (as it happens in applications of the Kolmogorov model).

Theorem 2.5 (Bayes formula). *For any S-sequence (collective) x and any subset O of the label set L_x, such that $\mathbf{P}_x(O) > 0$, the Bayes formula for conditional probability holds*:

$$\mathbf{P}(B|O) = \frac{\mathbf{P}_x(B \cap O)}{\mathbf{P}_x(O)}. \tag{2.17}$$

Proof. We have

$$\mathbf{P}_{z(O)}(B) = \sum_{s\in B\cap O} \mathbf{P}_x(s|O) = \sum_{s\in B\cap O} \mathbf{P}_x(s)/\mathbf{P}_x(O).$$

Thus in the von Mises model the Bayes formula for conditional probabilities is a theorem and not a definition; compare with the Kolmogorov model. By using the Bayes' formula (2.17) we obtain:

Theorem 2.6 (Formula of total probability). *Let x be an S-sequence and let $\{O_k\}$ be a partition of the label set L_x:*

$$L_x = \bigcup_k O_k \quad \textit{and} \quad O_k \cap O_l = \emptyset,$$

and $\mathbf{P}_x(O_k) > 0$ for any k. Then, for any $C \subset L_x$, we have:

$$\mathbf{P}_x(C) = \sum_{k=1}^{m} \mathbf{P}_x(O_k)\mathbf{P}_x(C|O_k). \tag{2.18}$$

2.3 Combining and Independence of Collectives

The material presented in this section will not be used in our contextual probabilistic considerations. We present it simply to give a more complete picture of the von Mises model. Therefore it is possible to read this section later. One might jump directly to Part II.

In the two basic operations, mixing and partition, discussed in Sect. 2.2, one single S-sequence (or collective) x served each time as the point of departure for the construction of a new S-sequence (collective). We now consider the problem of *combining* of two or more given S-sequences (collectives).

Let $x = (x_j)$ and $y = (y_j)$ be two S-sequences with label sets L_x and L_y, respectively. We define a new sequence

$$z = (z_j), \; z_j = (x_j, y_j). \tag{2.19}$$

In general such a z is not an S-sequence with respect to the label set $L_z = L_x \times L_y$.

Let $\beta \in L_x$ and $\alpha \in L_y$. Among the first N elements of z there are $n_N(\alpha; z)$ elements with the second component equal to α. As $n_N(\alpha; z) = n_N(\alpha; y)$ is a number of $y_j = \alpha$ among the first N elements of y, we obtain that

$$\lim_{N\to\infty} \frac{n_N(\alpha; z)}{N} = \mathbf{P}_y(\alpha).$$

Among these $n_N(\alpha; z)$ elements, there are a number, say $n_N(\beta|\alpha; z)$, of elements whose first component is equal to β.[5] The frequency $\nu_N(\beta, \alpha; z)$ of elements of the sequence z labeled (β, α) will then be equal to

$$\frac{n_N(\beta|\alpha; z)}{N} = \frac{n_N(\beta|\alpha; z)}{n_N(\alpha; z)} \frac{n_N(\alpha; z)}{N}.$$

We set

$$\nu_N(\beta|\alpha; z) = \frac{n_N(\beta|\alpha; z)}{n_N(\alpha; z)}.$$

Let us assume that:

For each $\alpha \in L_y$, the subsequence $x(\alpha)$ of x which is obtained by choosing x_j such that $y_j = \alpha$ is an S-sequence.[6]

Then, for each $\alpha \in L_y$, $\beta \in L_x$, there exists

$$\mathbf{P}_z(\beta|\alpha) = \lim_{N\to\infty} \nu_N(\beta|\alpha; z) = \lim_{N\to\infty} \nu_N(\beta; x(\alpha)) = \mathbf{P}_{x(\alpha)}(\beta).$$

The existence of $\mathbf{P}_z(\beta|\alpha)$ implies the existence of

$$\mathbf{P}_z(\beta, \alpha) = \lim_{N\to\infty} \nu_N(\beta, \alpha; z).$$

Moreover, we have

$$\mathbf{P}_z(\beta, \alpha) = \mathbf{P}_y(\alpha)\mathbf{P}_z(\beta|\alpha) \tag{2.20}$$

or

[5] We remark again that it is assumed that all data was collected in the sequence z and we need not perform new measurements to obtain all these numbers.

[6] In general such a choice of the subsequence $x(\alpha)$ of x is not a place selection.

$$\mathbf{P}_z(\beta|\alpha) = \mathbf{P}_z(\beta, \alpha)/\mathbf{P}_y(\alpha), \tag{2.21}$$

if $\mathbf{P}_y(\alpha) \neq 0$. Thus in this case the sequence z is an S-sequence with the probability distribution $\mathbf{P}_z(\beta, \alpha)$. This is the "joint probability distribution" of S-sequences of observations x and y. We can repeat all previous considerations for collectives (i.e., take into account the principle of randomness), see R. von Mises [311].

Definition 2.3. *S-sequences (collectives) x and y are said to be combinable if the sequence $z = (x, y)$ is S-sequence (collective).*

Definition 2.4. *Let x and y be combinable. Quantities $\mathbf{P}_z(\beta|\alpha)$ are called conditional probabilities.*

This is the definition of conditioning of one S-sequence (collective) with respect to another. It differs from the conditioning with respect to a subset O of the label set of a single S-sequence (collective), Sect. 2.2.

Definition 2.5. *Let x and y be S-sequences (collectives). The x is said to be independent from y if all $x(\alpha)$, $\alpha \in L_y$, have the same probability distribution which coincides with the probability distribution $\mathbf{P}_x$ of x.*[7]

This implies that

$$\mathbf{P}_z(\beta|\alpha) = \lim_{N\to\infty} \nu_N(\beta|\alpha; z) = \lim_{N\to\infty} \nu_N(\beta; x(\alpha)) = \mathbf{P}_x(\beta),$$

hence

$$\mathbf{P}_z(\beta, \alpha) = \mathbf{P}_y(\alpha)\mathbf{P}_x(\beta). \tag{2.22}$$

Thus the independence implies the factorization of the two-dimensional probability $\mathbf{P}_z(a, b)$.

[7] We recall again that the choice of a subsequence $x(\alpha)$ of x based on a label α for y is not a place selection in x. Thus in general there are no grounds for coincidence of probabilities $\mathbf{P}_{x(\alpha)}$ with the probability $\mathbf{P}_x$.