APPLIED MATHEMATICS SERIES

Edited by

I. S. SOKOLNIKOFF

INTRODUCTION TO THE THEORY OF PROBABILITY AND STATISTICS

APPLIED MATHEMATICS SERIES

THE APPLIED MATHEMATICS SERIES is devoted to books dealing with mathematical theories underlying physical and biological sciences, and with advanced mathematical techniques needed for solving problems of these sciences.

INTRODUCTION
TO THE THEORY OF
PROBABILITY
AND
STATISTICS

NIELS ARLEY
ASSISTANT PROFESSOR OF PHYSICS
INSTITUTE FOR THEORETICAL PHYSICS
UNIVERSITY OF COPENHAGEN

K. RANDER BUCH
ASSISTANT PROFESSOR OF MATHEMATICS
INSTITUTE FOR APPLIED MATHEMATICS
DENMARK INSTITUTE OF TECHNOLOGY

NEW YORK · JOHN WILEY & SONS, INC.
LONDON · CHAPMAN & HALL, LIMITED

FOURTH PRINTING, NOVEMBER, 1957

PRINTED IN THE UNITED STATES OF AMERICA

EDITORIAL PREFACE

The role which theories of probability and statistics have played in the development of physical and life sciences and social studies is too well known to require a justification for including this volume in a series of books devoted to applications of mathematics.

It is difficult to imagine a segment of mathematics that bears more intimately on everyday developments and which exerts more profound influence on modern scientific thought than the theories of probability and statistics. The unification of these theories on a level that imposes no demands on the real variable techniques, and yet preserves the spirit of modern developments, calls for extensive practical and theoretical knowledge as well as exceptional expository talents.

In the judgment of the editor, the authors of this volume have made a successful pioneering step by producing an eminently useful book. In its original edition in Danish it has already achieved an enviable record in continental Europe, and it is hoped that this augmented and revised version of it in English will render real service to applied mathematicians in the English-speaking world.

I. S. SOKOLNIKOFF

LOS ANGELES, CALIFORNIA
January, 1949

PREFACE

To the Danish Edition

The purpose of the present book is to give an elementary introduction to the theory of probability and statistics with special regard to its practical applications.

In recent years the theory of probability and statistics has undergone a very fruitful development with respect to the epistemological implications, the mathematical foundations, as well as the mathematical theory. We have tried to indicate in this book this modern development, all three directions of which are, in our opinion, of the greatest importance for all students of physics, chemistry, engineering and the other branches of modern science in which probability and statistics are used more and more as a working tool in the classification and analysis of our experiences.

The first three chapters are of a purely elementary nature. In the following chapters, we have presupposed a knowledge of the calculus, and, in the last chapter, a knowledge of the theory of linear equations. In the treatment of the theory of statistics we have followed the modern English school founded essentially by Fisher; and our discussion of the theory of errors makes use of the statistical methods developed principally by this school. These methods take into account the fact that, in practice, one often has to base his conclusions on a small number of observations. As is well known, the methods of the English school have been extremely fruitful in biology, medicine, agriculture, and other sciences. In the treatment of the theory of adjustment we have used the theory of matrices, principally because of the concise way in which it may be done. We believe that many of the more advanced students now have some knowledge of the matrix theory; nevertheless, we have included an appendix giving the most important definitions and theorems. Since the problems connected with the practical computations have been dealt with thoroughly in other textbooks, we have emphasized only the principal questions in our treatment.

We have chosen the examples from such different fields as ordnance, telephony, theoretical physics, and actuarial science. In so doing we hope that we have succeeded in showing the fundamental role of the theory of probability in modern science.

For convenience we have given the French and the German words

for the most important concepts; and, furthermore, we have added an extensive index, various statistical tables, and a list of some of the literature that is of special importance for the practical applications.

As for the symbols and notations, we have taken great care, first, to keep as closely as possible to those which seem to be already standardized, second, to carry out consistently the principle of sharply distinguishing between theoretical and empirical concepts by using Greek letters for the former and the corresponding Roman letters for the latter. Unfortunately, the field of probability and statistics shows a very confusing picture, a variety of symbols and notations for the same concepts still being in common use. It is to be hoped that an international standardization of at least the basic symbols and notations will be arrived at in a not too distant future.

We thank Professor H. Cramér of the University of Stockholm for having read the manuscript and for his kind permission to utilize, in various parts of the book, his treatment of the theory of probability. We also wish to thank Professor R. A. Fisher, and Dr. F. Yates, also Messrs. Oliver and Boyd, Limited, Edinburgh, for kind permission to reprint Tables II, III, V, and VI from their *Statistical Tables for Biological Medical, and Agricultural Research*. Finally, we shall highly appreciate criticism and comments from the readers, hoping to make use of them at a later date.

NIELS ARLEY
K. RANDER BUCH

Copenhagen, 1946

PREFACE

To the English Edition

Apart from minor alterations and additions, the present book follows the third Danish edition.

We should like to thank Mr. W. G. Stroud for his very great help in checking the translation, which was made by one of us (N.A.) during his stay as visiting assistant professor at Palmer Physical Laboratory, Princeton University, N. J., during the year 1946–1947. We thank Dr. G. U. Yule and Dr. M. G. Kendall, also Messrs. Charles Griffin and Co., Limited, London, for kind permission to reprint problems 87, 88, and 89 from *An Introduction to the Theory of Statistics*, twelfth edition, p. 309, Table 17.1, and p. 330, exercises 17.2 and 17.3.

NIELS ARLEY
K. RANDER BUCH

Copenhagen, 1949

CONTENTS

1.

THE CONCEPT OF PROBABILITY

§ **1.1.** Just as in geometry, mechanics, and other branches of applied mathematics, the purpose of the theory of probability is to give a mathematical description and analysis of a certain field of experience. The **theory of probability** is that part of mathematics which describes **statistical phenomena.**

In the treatment of problems from far different fields, both scientific and practical, one often meets events which appear in more or less *large numbers.* These events may be classified as follows:

1. The *same event* is *repeated* a certain number of times, the event being brought back to the same initial state before each repetition. Examples: repeated measurements of a physical quantity; a series of throws with a die.

2. The *same* event which changes with time is observed a certain number of *consecutive* times. Examples: the position of one Brownian particle is observed through a microscope at various times; the number of demands for a certain commodity in one shop on different days.

3. Several *distinct* events which in the problem at hand may be considered to be of the *same* kind are observed simultaneously. Examples: the positions of several Brownian particles are observed at the same time; the number of demands for a certain commodity in different shops on the same day; the heights of a number of soldiers; the yields of wheat from a number of different plots; the life times of a number of light bulbs.

4. The theoretical possibility that several *distinct* and *different* events are observed simultaneously. Examples: the measurement of the weights of a certain number of cows, hogs, and hens. This case, however, plays hardly any role in practice.

The characteristic feature of a statistical description is that, in contrast to a causal description, it is not applied to a single event but only to the whole course of a large number of events; i.e., it applies

Note: In this book, equations are numbered consecutively in each article and are referred to as follows: in §2.5, eq. 1 from the same article is referred to as (1); eq. 2 from §2.4 as (2.4.2); and so on.

Sections marked with an asterisk (*) are of a more difficult nature and may be omitted in a first reading.

only to properties belonging to a certain number of events considered as a whole. Thus, for example, in making a statistical study of childbirths, an investigator is interested only in how many children have been born, not in who has had a child. Also, in making such a statistical study one can investigate how the number of children is distributed between the two sexes, or among the single families, or among different groups of the population, and so on. One can also investigate how these facts change in time. Further, the averages of the numbers or their fluctuations can be emphasized.

However, the problem of applying a causal or statistical description in a specific instance is quite independent of whether the phenomenon in question is of a causal or statistical nature.

Example. Suppose that a person walks down a road beside which there are consecutively numbered telephone poles and observes the last digit on each pole. He can then describe these observations either by stating that the numbers come in a certain regular sequence, say 8, 9, 0, 1, 2, and so on (a causal description), or he can state how many times each of these numbers occurs in a large number of observations (a statistical description).

Which description, causal or statistical, will be used in a given case depends entirely on which is more appropriate. As a rule in both cases the description consists of simplifying and idealizing the phenomenon under consideration by constructing a certain model—in our case of a mathematical nature—which represents in the best way those features of the actual phenomenon which we, ourselves, regard as the most essential ones. Examples: the representation of observed phenomena in space by means of Euclidean geometry; the description of the motion of bodies by means of analytical mechanics; the application of simple mathematical laws to the description of statistical phenomena (cf. Chapter 10).

No such idealized model of a certain field of experience should be confused with reality itself. Therefore, any such model is in itself neither "true" nor "false," a condition that is further stressed by the fact that we can, as a rule, construct several different models representing the same phenomenon. *Whether or not it is legitimate to apply a certain model is determined entirely by the criterion that the model give a sufficiently satisfactory description of the real world* (cf. § 9.7).

§ 1.2. If a certain group of events is described in a statistical way, it is treated as a *random* phenomenon. This means that, either from principle or from practical reasons, it is considered impossible to predict

the "final state" of the phenomenon from the "initial state" and the known laws of nature.

If, on the contrary, it is regarded as possible to predict the final state, the description is a causal one. We wish to stress the fact that, since it may also be found that the final state of a statistically described phenomenon is always the same for the same initial state, *the statistical description contains the causal as a special limiting case.* However, this happens so rarely in practice that as a rule the statistical description gives a less idealized model of the real world than the causal one.

Random phenomena are characterized by the fact that they are conditioned by so many interacting causes that it is practically impossible to analyze them accurately enough to make exact predictions. The *first* reason for this random character may be that we cannot define the initial state accurately enough to determine the final state uniquely. This is the case:

(*a*) If a very small variation in the initial state can cause a large variation in the final state even though the phenomenon is simple in character. For example, an extremely small change in the initial rotational velocity of a roulette wheel can decide whether the final state is black or red. Thus, the principle of all *games of chance* is to construct them so that even a very small change in the initial state has a large effect on the final state.

(*b*) If the initial state is so complicated that it is practically impossible to ascertain it accurately enough to determine the final state uniquely, even though the phenomenon may again be very simple. For example, it is known that 1 gram of hydrogen contains 3×10^{23} molecules, but to measure all the coordinates and velocities of these molecules at a given time (the so-called *microscopic* state) is, of course, impossible. Therefore, to describe the initial state other quantities, such as thermodynamic pressure p, volume v, and absolute temperature T, have to be chosen; but, since such a *macroscopic* description does not determine the microscopic state uniquely, we cannot give a unique prediction of the final state from the macroscopic state alone. Consequently, in spite of the fact that the motion of each molecule is causally determined from the equations of classical mechanics, the state of this mass of hydrogen as a function of time must be considered a random phenomenon (cf. § 4.16).

Such a situation occurs whenever there is a contradiction between the two demands which must necessarily be made of the definition of the "state" of a phenomenon. The first of these demands is that the

properties which characterize the state allow all other properties in which we are interested to be determined uniquely by means of the laws of nature; the second, that the properties which characterize the state are directly observable and can be stated.

(*c*) If the initial state cannot in principle be measured because a measurement may produce an uncontrollable change in the phenomenon investigated. This is precisely the case for atomic phenomena, and thus quantum mechanics is decidedly different from classical physics in which it was assumed that the interaction between the measurement and the measured event could always be made arbitrarily small (cf. p. 5).

The *second* reason for the random character of a phenomenon may be related to the laws of nature involved. This is the case:

(*a*) If the relevant laws of nature are so complicated that in practice it is impossible to make the theoretically possible calculation of the final state. An example of this is a series of throws with a die; only a few throws show that the results are randomly distributed. Theoretically one could predict the result of a given throw if one knew the exact initial position and velocity of the die, its geometric form, its mass, its moments of inertia, the elastic properties of the die and of the table, and so on. Thus an extremely accurate analysis of all these cooperating causes would be necessary, but such an analysis would be so complicated that in practice it would be impossible.

(*b*) If the relevant laws of nature are not sufficiently well known; this is true, e.g., in most biological phenomena. Thus, if we measure the heights of a number of soldiers, the results are conditioned by a number of biological processes such as heredity and nutrition, the regularities of which are only partly known.

The *third* reason for the random character of a phenomenon is that all laws of nature are strictly valid only for *idealized* phenomena. As already mentioned, to deduce the general laws we have to *simplify* and *idealize* the phenomena, intentionally neglecting many factors and considering only one of the acting causes. An example is the law of inertia, in which friction is usually neglected. In reality the phenomena are always complicated and, furthermore, are subjected to disturbing factors such as changes in temperature and pressure, shocks, backlash in screws, and mechanical, electric, and magnetic actions from the surroundings and often from the measuring apparatus itself.

Thus, if we measure an electric current by means of an ammeter, the current is changed, since the meter itself has a certain resistance. In this example the disturbing factor may be calculated and the measured value corrected. However, this example is exceptional. Especially conspicuous examples of the fact that the measurement itself may affect the event considerably are encountered in biology and psychology. Thus an investigation of the function of a vital organ of a living animal may easily disturb the whole organism to such an extent that the animal dies. Similarly, if we wish to analyze an emotion such as fury in ourselves or in others, the investigation may easily result in the disappearance of the emotion. Such disturbing factors are often of incalculable magnitude and, as mentioned before for atomic phenomena, may in principle be uncontrollable. Nevertheless, these disturbing factors are just as decisive as the other causes in the determination of the phenomena. Hence every physical measurement is to a larger or smaller degree a random event giving different results in repeated measurements.

As a rule all three of these groups of factors appear simultaneously. Consider the throw of a die as an example. First, the initial state enters in a critical way since a very small change in the initial state may be decisive in giving a result of 2 or 6. Second, the laws of nature are so complicated here that we could not carry through an accurate calculation even if we knew the initial state exactly, and, finally, there occur disturbing factors such as air resistance. Another well-known example of a random phenomenon in which all three groups of factors enter is found in ordnance, in which the points of impact of a series of shots will always be different.

§ 1.3. An *observation* is defined as the statement of a definite result of a random phenomenon. As a rule the result of an observation will consist of one or more numbers, but it may be characterized in other ways. In the example of the throws of a die, the observation consists of throwing the die and reading the number. The result of the observation is one of the numbers 1, 2, 3, 4, 5, or 6. In considering the motion of a molecule, the observation consists of measuring the position and velocity of the molecule at a certain time, and the result consists of six numbers, viz., the three coordinates and the three components of the velocity, each of which six numbers may assume any value between $-\infty$ and ∞. On the other hand, if we consider an investigation within population statistics, the result of an observation need not be a number. Thus, if the observation consists of reading from the census record the sex of an individual, the result is either "male" or "female."

§1.4. Let us now consider a definite random phenomenon in which we have made a series of n observations—in one of the four ways discussed in §1.1. Let us count the number of results, n_A, belonging to a certain specific class. Here we say that the *event* A has happened. The number n_A is called the **absolute frequency** of A, and the fraction

$$f(A) = \frac{n_A}{n}$$

is called the **relative frequency** in the given series of observations. *Experience* shows that the random variations or statistical fluctuations, as they are also called, are, as a rule, smoothed out, the smoothing process following a certain law which we will call the **random law**,[1] the validity of which is a fundamental condition for all applications of the theory of probability to the description of real phenomena:

If, in a definite, accurately stated category $\mathfrak{C}$ of observations, one calculates the relative frequencies of a certain event A in different series of observations, experience shows that the numbers so obtained deviate very little from each other if each series consists of a very large number of observations.

In other words, if we plot these relative frequencies on a line, they are grouped about a common value—a sort of "cluster point," the spread of the cluster decreasing with increasing number of observations in each series. To denote this form of randomness the word **stochastic randomness** has been introduced.

TABLE 1

Number of Units per Group

25	250	2500
Number of Units Rejected per Group and Corresponding Percentages		
1 (4)	12 (4.8)	157 (6.28)
4 (16)	14 (5.6)	152 (6.08)
0 (0)	17 (6.8)	157 (6.28)
0 (0)	11 (4.4)	136 (5.44)
1 (4)	22 (8.8)	152 (6.08)
1 (4)	9 (3.6)	135 (5.40)
2 (8)	15 (6.0)	143 (5.72)
0 (0)	14 (5.6)	160 (6.40)
1 (4)	21 (8.4)	149 (5.96)
1 (4)	8 (3.2)	153 (6.12)

[1] This law of experience is often called the **law of large numbers,** but this name is also often used for a *mathematical* law (cf. (8.1.14)). This dual use of the name is unfortunate and has resulted in much confusion in the historic development of probability.

Example 1. As an example of stochastic randomness we shall consider the percentage of rejects in the manufacture of an industrial product. In Table 1 we give the number of rejected units in each of 10 sampled groups, each group containing first 25, then 250, and finally 2500 units respectively. The figures in parentheses are the respective percentages. The variation in the percentage of rejects is illustrated in Fig. 1, in which the percentages found are plotted on a straight line. Each of the values found is indicated by a dot over the corresponding point on the line. It should be noted that the percentages of rejects found (and thus the relative frequencies) are grouped around a common value of about 6% and that the spread of the variations is smaller when a group contains more units.

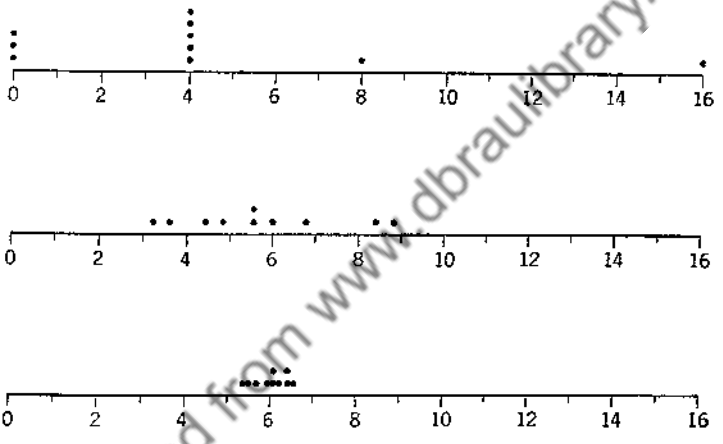

FIG. 1.

Example 2. In practice, forms of randomness other than stochastic are also encountered. Let us again consider the example, § 1.1, and let the observation of the number 2 as the last number on the telephone pole be the event A. In general, $n_A/n \sim 1/10$. If, however, we count only every tenth pole, we will find that $n_A/n = 1$ for all n's if we start counting at a pole with the number 2 and that $n_A/n = 0$, otherwise.

In this extreme case the relative frequencies are not grouped around a common value. Consequently, it is inappropriate to use a statistical description—a fact which is also obvious since a causal description can be given. For a "real" randomness we shall expect that n_A/n is grouped around a common value even if we make a selection of our observations in a "random" way—as would be the case if the numbers had been obtained, not by this method but by choosing the

last digit of the numbers on a certain page of a telephone directory. We shall leave it to the reader to try this for himself.

If experience shows that the random phenomenon considered is a stochastically random phenomenon, there is "something," i.e., some physical property of the phenomenon considered, which appears to be *constant*. For a mathematical description it is now an obvious procedure to idealize this experience by abstracting from the deviations between the relative frequencies and to introduce one definite number to represent the "true" value of the relative frequencies of the event in the same way that we introduce one definite number to represent the "true" value of any other physical quantity (cf. § 11.2). Thus we interpret the relative frequencies as the experimental values of one and the same physical constant which is determined by the nature of the event A and of the category of observations $\mathfrak{C}$ (cf. § 1.6). But, as already stressed, the legitimacy of this interpretation is based only on the fact that the interpretation leads to an appropriate description of the statistical phenomena that are met in practice.

This physical constant, for which the relative frequencies are experimental values, is called the **probability** *of the event A in the category of observations $\mathfrak{C}$ and is denoted by $P(A)$.* Because of the manner in which $P(A)$ has been introduced, it is also called the **true relative frequency** of A.

We wish to stress that, with this definition, the probability is *not* defined for an event which can occur only once, i.e., for an event which cannot be reproduced. For example, we cannot speak of the probability of the event that a certain tennis player will win a specific match in spite of the fact that in everyday language the word probability is applied to such events.[1]

Thus, this number, the probability, is determined by A and $\mathfrak{C}$ in the same sense as the weight of an iron bar is determined by the specific iron bar even though repeated measurements of this weight give more or less different values. In the same way that experience shows that, as a rule, the accuracy of a measurement can be increased by applying more sensitive measuring instruments, it also, as a rule, shows that the accuracy of the empirical determination of a probability can be increased by making a larger number of observations

[1] Certain authors have also tried to define the mathematical probability of non-reproducible events as the degree of reasonable belief; but then the difficulties are how to obtain the numerical values in concrete cases and how to compare given numerical values with experiences. The best-known exponents of this school are J. M. Keynes, *A Treatise on Probability*, London, 1921; and H. Jeffreys, *Theory of Probability*, London, 2 ed., 1948.

(cf. Example 1). But, in the same way that the accuracy of any physical measurement cannot be increased indefinitely, the accuracy of a measurement of a probability cannot be increased without limit. Every empirical number can almost always be determined only within a certain limited accuracy, and therefore in practice it is meaningless to speak of performing a limiting process by making the number of observations "infinitely large" (cf. p. 11).

The introduction of the concept of probability is obviously analogous to the introduction of every other idealized concept by means of which we describe and classify observed phenomena (e.g., Euclidean straight line in geometry, position as a continuous function of time in mechanics). In the same way that we always unconsciously make the relevant idealizations, as for instance from "observed straight line" to "Euclidean straight line," and in the same way that we always try to keep the mathematical model separated from the corresponding reality, *in probability we must now accustom ourselves partly to the unconscious idealization from the relative frequency of an event to its probability and partly to the separation in our minds of the two concepts, frequency and probability.* In the following we shall use, as far as possible, different types of letters for the two concepts: Roman letters for all concepts referring to reality, corresponding Greek letters for the related concepts in our model.

§ 1.5. In certain problems it is possible to deduce theoretically the probabilities of the events considered from certain simple hypotheses about the nature of the phenomena. Such probabilities are called **a priori** probabilities, i.e., determined in advance. In contrast, probabilities determined from a measurement of the corresponding relative frequencies are called **a posteriori** probabilities, i.e., determined afterward.

A priori probabilities are found especially in theoretical physics, e.g., in kinetic theory of gases and in quantum theory. Also a priori probabilities are met with in the theory of games of chance. This application of probability is of especial historic interest since the treatment of such problems led to the rise of the theory of probability in the seventeenth century, in particular, from the investigations of the French mathematicians Fermat and Pascal. If, for example, we play with a die, we shall expect from reasons of symmetry that each side will come up as frequently as any other in a long series of throws. Since in all there are six possible results of a single throw, it follows that the relative frequency of each of the six single results must lie very close to 1/6; therefore it is a natural hypothesis to ascribe the probability 1/6 to each of the six results. Correspondingly, we must ascribe the

probability 1/52 to the event of selecting a queen of hearts from a deck of cards and the probability 28/52 to the event of selecting a card that is either red or a king, or both. This number 28/52 is obtained by noting that the event in question has occurred only if the selected card is either one of the 26 red cards or one of the 2 black kings. However, it should be remarked that in all such a priori determinations of probability we make certain, usually tacit, hypotheses, e.g., that the die is not "false," that we play "honestly," that the cards are well "shuffled," and so on.

In all, the problems of games of chance mentioned above have two common features: (*a*) that there are only a finite number of possible results; and (*b*) that from our knowledge of the play we, "a priori," regard certain results as "equally likely," i.e., equally probable.[1] The **classical definition of probability** fits these facts very closely by simply defining the probability of an event as the ratio of the number of results favorable to the event to the total number of possible (and equally likely) results. In addition to the fact that this definition is logically a circle definition, since "equally likely" can be defined only as "equally probable," which was to be defined, it can be applied only when the observation in question can give a finite number of different results. Furthermore, there is often doubt as to what has to be understood by the phrase "equally likely results," as shown by the following example.

Example. Assume that we have three identical bureaus, A, B, and C, each with two drawers. In A, each drawer contains a gold coin; in C, each drawer a silver coin; and, in B, one drawer contains a gold coin, the other, a silver coin. We choose at random a bureau, open one of the drawers, and find a gold coin. What is the probability that the bureau chosen is just B? First, it might be argued that, since only A and B contain gold coins, the number of possibilities is 2 and only 1 of these is favorable. Consequently, the probability is 1/2. But it can also be argued that, in all, there are 3 gold coins of which A contains 2 and B only 1. The number of possibilities is 3, of which only 1 is favorable, and thus the probability is 1/3. Which of the two arguments is correct? If the opened drawer had contained a silver coin, we should also have obtained the answers 1/2 and 1/3 in the first and second arguments, respectively. Hence the probability

[1] We stress that it is our positive knowledge, e.g., of the physical properties of a die, which leads us to expect each result as equally probable. Nevertheless it is sometimes stated that this expectation is due to our ignorance of the behavior, e.g., of a die. Cf. the discussion of this point by H. Poincaré, as given in *Handbuch der Physik*, Vol. IV, p. 67, Berlin 1929.

of the chosen bureau being B is independent of whether or not we open a drawer. However, if we do not open a drawer, the three bureaus are, of course, equally likely and the required probability is 1/3. Thus the first argument leads to a contradiction and has to be rejected (cf. Example 2, §3.8).

This classical definition of probability, due essentially to Laplace, was very appropriate in the infancy of probability when the theory was applied only to the description of games of chance. However, as we have remarked, this definition is much too narrow for the description of general statistical phenomena, since, first, there may be more than a finite number of possible results, and, second, it may not be possible to reduce the phenomenon considered to a number of equally likely possibilities. For example, from the classical definition it would be impossible to define the probability of the result 6 with a loaded die—quite apart from the fact that we can also play falsely with a true die. From the classical definition we often cannot speak of probability in a certain event. For example, in the case of the probability of a boy's being born, what should be understood by equally likely results? Furthermore, the classical definition always gives a rational number, so that the extension to a more general definition of probability, where a probability may be an irrational number, is to a certain extent analogous to the extension of the concept of numbers from rational to real numbers.

Finally, various authors have tried to introduce the concept of probability in a way different from ours, which is essentially due to Fréchet, and from the classical. The best-known exponent of this school, von Mises, defines the probability of an event A as the limit of the frequency, n_A/n, for $n \to \infty$, in one definite infinite series of observations (called by him a *collective*). The existence of this limit is postulated as the first axiom of the theory. A definition of this type seems at first sight very attractive, but a closer analysis has shown it to lead into great mathematical difficulties.[1] Apart from these difficulties, such a constructive definition involves a mixture of reality and model which has been abandoned in all other branches of applied mathematics in favor of an axiomatic treatment. For example, we never introduce the concept of a Euclidean straight line as the limit for infinitely decreasing width of a line drawn on a blackboard with a piece of chalk, or the concept of a mass point as the limit for infinitely decreasing radius of a real body.

[1] Cf. Fréchet, "The Diverse Definitions of Probability," and: "Exposé et discussion de quelques recherches récentes sur les fondements du calcul des probabilités."

§1.6. We wish to stress the fact that by the definition of probability (§1.4) we must not consider only the nature of the event but also the nature of the category of observations. This latter is, as a rule, sufficiently defined in the formulation of the problem itself. But often the necessity of exactly defining how the observation is to be performed has been overlooked, with the result that wrong conclusions and paradoxes are arrived at. If we ask for the probability of a hit for a given gun and a given target, there can be no doubt about the operations that have to be performed. However, if the probability of a person's dying within one year after purchase of an insurance policy is required, the answer is much more difficult to determine. Within which group of men must we calculate the relative frequency of death? Obviously, different values for the probability are obtained, dependent on whether the group of men of the same age is chosen or the group consisting only of males of the same age or, furthermore, whether the vocation, the state of health, and so on, are taken into account.

This fact is shown even more conspicuously in the so-called "Bertrand's paradox." Here the observation consists of randomly drawing a chord in a circle of radius r and reading off its length. Suppose that the probability that the length of the chord is smaller than the side of the inscribed equilateral triangle, that is, smaller than $\sqrt{3}\,r$, is required. Because the process of "randomly drawing a chord" is not a well-defined operation, it is possible to give many contradictory answers. We shall give only the two following solutions:

1. The perpendicular distance from the center of the circle to the chord is a number between 0 and r. The chord is smaller than $\sqrt{3}\,r$ if this distance is larger than $r/2$. If the probability is measured as the ratio of the favorable length $r/2$ to the possible length r, the result is $1/2$.

2. The central angle of the chord is a number between 0 and π. The chord is smaller than $\sqrt{3}\,r$ if the central angle is smaller than $2\pi/3$. If the probability is measured as the ratio of the favorable central angle $2\pi/3$ to the possible angle π, the result is $2/3$.

2.

THE FOUNDATIONS OF THE THEORY OF PROBABILITY

§ **2.1.** Having discussed the concepts which experience shows appropriate to the description of statistical phenomena, we shall now state **the fundamental laws** or **axioms,** that is, the relations among the numbers we introduce as probabilities, which must hold true if our theory is to give us a suitable description of the real world. For the probability of a single event we may, as a rule, choose freely among infinitely many numbers (within a certain interval, as, e.g., between 5.8% and 6.3% in Example 1, § 1.4). However, by the simultaneous introduction of probabilities for several events, we cannot introduce these as we please because there exist certain relations between the corresponding relative frequencies and because from our definition of probability (§ 1.4) we must demand that the same relations hold true for the probabilities. Thus we must lay down certain fundamental laws, which will form the foundations of all the following deductions, but which cannot themselves be proved mathematically. (This is analogous, e.g., to the axioms of Euclidean geometry, Newton's laws in mechanics, and Maxwell's equations in electrodynamics.)

Since a relative frequency is always of the form $f = n_A/n$, where $0 \leqq n_A \leqq n$, we have $0 \leqq f \leqq 1$, and therefore we must demand the same relation for any probability.

First axiom: *The value of a probability, P, is a number between* 0 *and* 1, *both limits included:*

$$0 \leqq P \leqq 1. \tag{I}$$

§ **2.2.** By the phrase a **certain event,** denoted in the following by E, we understand an event which occurs in every observation. The relative frequency, $f(E)$, of a *certain* event is thus always

$$f(E) = \frac{n}{n} \equiv 1,$$

and therefore we must demand the same for its probability, $P(E)$.

Second axiom: *The probability of a certain event is* 1:

$$P(E) = 1. \tag{II}$$

We stress the fact that the converse need not be true. If an event has the probability 1 (denoted as a **stochastically certain** event), this need not mean that the event is certain in the usual sense but only that it is **practically certain,** that is, that its relative frequency must be expected to be very close to 1 if the number of observations is large (cf. §9.2). In the following, when we speak of a *certain* event we shall mean a stochastically certain event without explicitly stating every time that this usage of the language deviates slightly from the commonplace.

§2.3. By the phrase an **impossible event,** denoted in the following by O, we understand an event which cannot occur in any observation. The relative frequency, $f(O)$, of an *impossible* event is thus always

$$f(O) = \frac{0}{n} \equiv 0,$$

and therefore we must demand the same for its probability, $P(O)$.

Third axiom: *The probability of an impossible event is* 0:

$$P(O) = 0. \tag{III}$$

Here, as in §2.2, the converse need not be true. If an event has the probability 0 (denoted as a **stochastically impossible** event), we can conclude only that the event is **practically impossible,** that is, that its relative frequency must be expected to lie very close to 0 if the number of observations is large (cf. §9.2). In the following, when we speak of an impossible event, we shall mean a stochastically impossible event without explicitly stating every time that this usage of the language also deviates slightly from the commonplace. In the last two topics we have touched upon one of those points in which one must remember not to confuse reality with the model, for a theory can prove only statements regarding the model, not reality.

§2.4. Let us consider two events A and B and a series of n observations. In each observation one and only one of these four possibilities may occur:

1. A has occurred, but not B.
2. B has occurred, but not A.
3. Both A and B have occurred.
4. Neither A nor B has occurred.

Let n_1, n_2, n_3, and n_4 denote the number of times in the n observations

that the respective possibilities have occurred, so that

$$n_1 + n_2 + n_3 + n_4 = n. \tag{1}$$

Now we can form the following relative frequencies:

Relative frequency of A (independent of B):

$$f(A) = \frac{n_1 + n_3}{n}. \tag{2}$$

Relative frequency of B (independent of A):

$$f(B) = \frac{n_2 + n_3}{n}. \tag{3}$$

Relative frequency of either A or B or both (which event we shall denote by $A + B$):

$$f(A + B) = \frac{n_1 + n_2 + n_3}{n}. \tag{4}$$

Relative frequency of both A and B (which event we shall denote by AB):[1]

$$f(AB) = \frac{n_3}{n}. \tag{5}$$

Relative frequency of B under the condition that A has occurred (which event we shall denote by $B|A$):

$$f(B|A) = \frac{n_3}{n_1 + n_3}. \tag{6}$$

Relative frequency of A under the condition that B has occurred:

$$f(A|B) = \frac{n_3}{n_2 + n_3}. \tag{7}$$

§2.5. For the six quantities given in (2.4.2) to (2.4.7) we have the following relations:

$$f(A + B) = f(A) + f(B) - f(AB) \tag{1}$$

$$f(AB) = f(A)f(B|A) = f(B)f(A|B). \tag{2}$$

Therefore, as discussed p. 13, we must demand that the same relations hold true for the corresponding probabilities. Thus, from (1), we have:

[1] The reason for using the symbols $A + B$ and AB in this connection is that the usual algebraic rules apply to them to a large extent. Verify this.

Fourth axiom: *The probability that at least one of two events occurs is equal to the sum of the probabilities of each event minus the probability of both events occurring simultaneously:*

$$P(A + B) = P(A) + P(B) - P(AB). \tag{IV}$$

This axiom is called the **addition law of probability** or, often, the law of **"either, or."**

Example. A card is drawn from each of two separate, well-shuffled decks of cards. The probability that either one or the other is the queen of hearts is then

$$1/52 + 1/52 - 1/52 \cdot 52 = 103/2704.$$

Here the third term was calculated as follows: calculating the probability for both cards being queen of hearts we have $52 \cdot 52$ possibilities, this being the number of ways in which we can match one card from one deck with one card from the other. However, of these $52 \cdot 52$ possibilities, only 1 gives the result: queen of hearts-queen of hearts.

§2.6. The probability $P(B|A)$ which corresponds to the relative frequency $f(B|A)$ defined in (2.4.6) is called the **conditioned** or **relative probability** of B under the condition A. Thus $P(B|A)$ gives the probability that B occurs on the condition that A has already occurred. To distinguish it from $P(B|A)$, $P(B)$ is called the **absolute probability** of B. $P(A|B)$ is defined in the corresponding way. From (2.5.2) we now see that the last axiom we have to lay down is the following:

Fifth axiom: *The probability of both of two events occurring is equal to the product of the absolute probability of one event and the conditioned probability of the other:*

$$\begin{aligned} P(AB) &= P(A)P(B|A) \\ &= P(B)P(A|B). \end{aligned} \tag{V}$$

This axiom is called the **multiplication law** of probability or, often, the law of **"both, and."** It is also called the law of **compound probabilities.**

Example. We draw successively 2 cards from the same deck, without replacing either card. The probability of both cards being hearts is given by $13/52 \cdot 12/51 = 1/17$. Having found the first card to be a heart, we have only 51 cards left in the deck, of which only 12 are hearts. Thus the conditioned probability for the second card's being a heart is 12/51.

§ 2.7. It is the purpose of mathematical probability to make deductions from the five fundamental laws, axioms I–V. The probabilities occurring in the equations obtained by this mathematical deduction are only arbitrary numbers which satisfy the axioms independently of the manner in which they have been obtained. However, in making practical applications of the theory, whenever the word "probability" is used, it is necessary to think of the way in which it is measured in practice, i.e., as the number about which the relative frequencies of the events considered are grouped—the "true" value of these relative frequencies. In this connection it is often a convenient picture to think of the class of each and every result that the observation considered can give and then to think of the probability as the "relative frequency" of the event in this imagined, infinite class which is denoted by such words as *population*, *ensemble*, *assemblage*, *universe*, and *collective*.[1] Since, in general, these "relative frequencies" will be of the form ∞ / ∞, it is impossible, without further comment, to give such a picture a well-defined meaning, but we must think of it as a sort of shorthand expression for the probabilities we have introduced.[2] When this fact is remembered, such a picture may often be of great value both heuristically and mnemotechnically, especially in statistics.

Every problem in probability will now be of the following form: one starts from certain probabilities $P_1, P_2, \cdots$, which are given from certain theoretical considerations or hypotheses, i.e., a priori, or given from experience by the corresponding relative frequencies, $f_1, f_2, \cdots$, i.e., a posteriori, or, finally, given as arbitrary constants, the values of which have to be determined later. Such constants are called **parameters.** Next, by means of the mathematical theorems deduced from axioms I–V, certain other probabilities, $P_1', P_2', \cdots$, are calculated as functions of $P_1, P_2, \cdots$. The given values of $P_1, P_2, \cdots$ are inserted in these functions, and finally the verification of the theory consists of comparing the values thus obtained with the corresponding relative frequencies, $f_1', f_2', \cdots$, the latter being the experimental values of $P_1', P_2', \cdots$ (cf. Chapter 9). If, on the other hand, $P_1, P_2, \cdots$ are arbitrary parameters or depend on such, we

[1] We note that unfortunately these same words are also used for a finite class of observational results, thereby confusing reality with our model of it (cf. § 1.4). We therefore ought always to speak of either an empirical or a theoretical population, ensemble, and so forth.

[2] In higher mathematics, i.e., in the so-called theory of abstract measure, this picture can be given a well-defined meaning. By this, probability can be formulated in a very general way, as first done systematically by A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*.

try to choose their numerical values in such a way that agreement between the theoretical and the experimental values of $P_1', P_2', \cdots$ is the best possible.

We wish to stress the two following facts. First, the theory of probability must always start from certain probabilities as given in the problem considered—just as the initial positions and velocities of the bodies whose motions are described in mechanics. Second, the theory of probability will always give results in the form of probabilities. Consequently, the theory of probability by its very nature can never teach us anything about the actual course of a single event.

In Chapters 3 to 8 we shall present the purely mathematical theory, returning in Chapter 9 to the relation between theory and experience; and, finally, in Chapters 10–12 we shall give various practical applications of the theory.

3.

ELEMENTARY THEOREMS

§ **3.1.** Under certain assumptions the addition law of probability, axiom IV, reduces to a simpler form. If we assume that the events A and B **exclude each other** (i.e., cannot occur simultaneously, which fact we symbolize by $AB = 0$), we have from axiom III, $P(AB) = 0$, and thus axiom IV becomes

$$P(A + B) = P(A) + P(B). \tag{1}$$

If two events exclude each other, the probability of either the one or the other occurring is equal to the sum of the probabilities of each occurring separately.

Example. In a throw with a die the probability of getting either 3 or 4 is

$$1/6 + 1/6 = 1/3.$$

§ **3.2.** If $P(B|A) = P(B)$, i.e., if the probability of B is independent of whether or not A has occurred, we say that the event B is **stochastically independent** of the event A. From axiom V, assuming $P(B)$ different from 0, we then find that $P(A|B) = P(A)$, i.e., that A is also stochastically independent of B. Therefore we may say that A and B are stochastically independent. Here axiom V becomes

$$P(AB) = P(A)P(B). \tag{1}$$

If two events are stochastically independent, the probability of both occurring is equal to the product of the probabilities of each event occurring separately.

Example. One card is drawn from each of two decks. The probability that both cards are hearts is then

$$13/52 \cdot 13/52 = 1/16$$

(cf. the example, § 2.6).

When applying the theorems (3.1.1) and (3.2.1), we must always remember that they hold true *only* under the conditions stated. How-

ever, this is often likely to be forgotten, and false conclusions thereby are reached. In such simple applications to games of chance as are used here as examples, it is, as a rule, easy to decide whether or not the conditions mentioned are satisfied. However, in less simple applications it is often more difficult to make this decision, and for them it is therefore safer to start from axioms IV and V themselves.

We wish to stress that the concepts *causally independent* and *stochastically independent* are not synonymous, although in practice they are often thought of as being identical, assuming that, if the events are causally independent, then $P(B|A) = P(A)$, i.e., that the events are also stochastically independent, and vice versa. In the following we shall omit for shortness the word stochastically and simply say that two events are independent when we mean that they are stochastically independent. Thus, in practice, when we assume that two events are independent, it is, as a rule, a hypothesis, the legitimacy of which can be decided only by experiments.

§3.3. The theorems (3.1.1) and (3.2.1) may easily be extended to more than two events. The event that at least one of ν events, A_1, A_2, $\cdots$, A_ν, occurs is symbolized by $A_1 + A_2 + \cdots + A_\nu$. Let the ν events exclude each other two-by-two, which fact we symbolize by writing $A_iA_k = O$, $i \neq k$. Under these conditions we obtain from (3.1.1)

$$P(A_1 + A_2 + \cdots + A_\nu) = P(A_1) + P(A_2) + \cdots + P(A_\nu). \quad (1)$$

Exercise. Verify this.

For a general theory it has turned out to be appropriate to generalize (1) to the case where we have an infinite number of events which exclude each other two-by-two, although such a case can, of course, never be encountered in practice. Since this generalized form of (1) does not follow from (1) we must lay it down as a special, independent axiom which is called the *axiom of complete additivity*.

Sixth axiom: *If A_1, A_2, $\cdots$ is an arbitrary infinite series of events which exclude each other two-by-two, i.e., $A_iA_k = O$, $i \neq k$, we have*

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad \text{(VI)}$$

§3.4. The event that ν events A_1, A_2, $\cdots$, A_ν occur simultaneously is symbolized by $A_1A_2 \cdots A_\nu$. If every one of the ν events is independent of each combination of all the other events we obtain from (3.2.1):

$$P(A_1A_2 \cdots A_\nu) = P(A_1)P(A_2) \cdots P(A_\nu). \quad (1)$$

Exercise. Verify this.

§3.5. If it is *certain* that at least one of the ν events $A_1, A_2, \cdots, A_\nu$ occurs, we symbolize this fact by writing

$$A_1 + A_2 + \cdots + A_\nu = E. \tag{1}$$

From axiom II we then have

$$P(A_1 + A_2 + \cdots + A_\nu) = 1. \tag{2}$$

Furthermore, if the ν events exclude each other two-by-two, i.e., if $A_i A_k = O$, $i \neq k$, we obtain from (3.3.1) and (2)

$$P(A_1) + P(A_2) + \cdots + P(A_\nu) = 1. \tag{3}$$

This important equation is often useful as a control for the correctness of a calculation and should be applied whenever possible. (See, e.g., §3.7.)

By the symbol $\bar{A}$ we denote the event that A does *not* occur. The event $\bar{A}$ is called the **opposite** or **complementary** event of A. Since at least one of the two events A and $\bar{A}$ must occur, and since they cannot occur simultaneously, we have in our symbols, $A + \bar{A} = E$ and $A\bar{A} = O$. From (3) we then have

$$P(A) = 1 - P(\bar{A}). \tag{4}$$

§3.6. It is often a useful device to calculate the probability of the complementary event $\bar{A}$ instead of the probability of the event A, itself. The following problem is a historically famous example.

In 1654 the French gambler De Méré stated in a letter to Pascal that he had made an observation, the result of which surprised him. Throwing 1 die 4 times, he found that the probability of obtaining the result 6 at least once turned out to be larger than the probability of obtaining the result double 6 at least once in 24 throws with 2 dice. In throwing the single die there are 6 possible results of which 1 is favorable; in throwing 2 dice there are 6 times as many possible results of which again only 1 is favorable. De Méré therefore argued that it must be necessary to make 6 times as many throws with 2 dice as with 1 die in order to obtain the same probability of a favorable result. However, Pascal showed that this argument is wrong; in the first game we have to calculate the probability, P_1, of obtaining at least once the result 6 in 4 throws with 1 die. The complementary of this event is that the result 6 is not obtained in any of the 4 throws. These 4 throws being independent, we obtain from (3.4.1) the result that the probability of this event is $(5/6)^4$, since the probability of not obtaining

the result 6 in 1 throw is 5/6. Consequently, we find from (3.5.4)

$$P_1 = 1 - (5/6)^4 = 0.518.$$

In the same way, Pascal found that the probability in the second game is

$$P_2 = 1 - (35/36)^{24} = 0.491.$$

These results were in agreement with the observation of De Méré that P_1 is slightly larger than $1/2$, P_2 slightly less than $1/2$.

§3.7. Bernoulli's problem. Let us consider an event A with the probability θ. We perform ν observations and ask for the probability P_r that A occurs in just r of these observations. First, let us assume that A occurs in the first r of the ν observations and not in the rest. From (3.4.1) the probability for this event is

$$\theta^r(1 - \theta)^{\nu - r}. \tag{1}$$

However, this probability is independent of where among the ν observations the favorable event occurs. From (3.3.1) we then obtain the desired probability P_r by multiplying (1) by the number of ways in which we can select r elements out of ν

$$ {}_\nu C_r = \binom{\nu}{r} = \frac{\nu(\nu - 1) \cdots (\nu - r + 1)}{1 \cdot 2 \cdots r} = \frac{\nu!}{r!(\nu - r)!}.$$

Consequently, we have

$$P_r = \binom{\nu}{r} \theta^r(1 - \theta)^{\nu - r}. \tag{2}$$

This formula is called the **binomial law.**[1] It gives the probability of the event A occurring just r times among ν observations. From (3.3.1) we then have that the probability, $P_{\geq r}$, of A occurring at least r times is given by

$$P_{\geq r} = P_r + P_{r+1} + \cdots + P_\nu = \sum_{i=r}^{\nu} P_i = \sum_{i=r}^{\nu} \binom{\nu}{i} \theta^i(1 - \theta)^{\nu - i}. \tag{3}$$

[1] In English literature ν, θ, and $1 - \theta$ are often denoted n, p, and q respectively. However, being parameters, they must, in our terminology, be denoted by Greek letters.

Putting $r = 0$ into (3), we find by means of the binomial formula

$$P_{\geq 0} = \sum_{i=0}^{\nu} \binom{\nu}{i} \theta^i (1-\theta)^{\nu-i} = [\theta + (1-\theta)]^\nu = 1. \tag{4}$$

Since r can assume only one of the values 0, 1, 2, $\cdots$, ν, (4) is in agreement with (3.5.3).

***Example.** Formula (1) has an important physical application. Let $\lambda\, dt$ denote the probability of a radioactive atom decaying in the "infinitely small" time dt. What is the probability of the atom's being "alive," i.e., not yet decayed, at the time t? We assume that the atom is created at the time $t = 0$, divide the time interval from 0 to t into ν equal parts each of length $\Delta t = t/\nu$, and apply (1), putting $r = 0$ and $\theta = \lambda\, \Delta t$. In the limit $\nu \to \infty$ we obtain the desired probability

$$\lim_{\nu \to \infty} (1 - \lambda\, \Delta t)^\nu = e^{-\lambda t}, \tag{5}$$

since $\lim_{x \to 0} (1+x)^{1/x} = e = 2.718 \cdots$. The time interval for which the probability of decay is 1/2 is called the half-life and from (5) is given by $T = \ln 2/\lambda$.

Exercise 1. A battery of guns, A, has for each shot a probability of hit 3/5; and another battery of guns, B, has a probability of hit 1/2. A fires 2 shots, and B 3 shots. Find the probabilities of exactly 0, 1, 2, 3, 4, or 5 hits. Check that their sum is equal to 1.

Exercise 2. Apply (3) to the solution of De Méré's problem §3.6.

***§3.8. Bayes' theorem.** In axiom V we may also think of $P(A)$ and $P(AB)$ as being the known quantities; then V may be written

$$P(B|A) = \frac{P(AB)}{P(A)}, \tag{1}$$

assuming $P(A) \neq 0$.

Example 1. The probability of an atom's still being alive at the time $t + t'$ (event B) under the condition that it is alive at the time t (event A) is given by (1) and (3.7.5)

$$P(B|A) = \frac{e^{-\lambda(t+t')}}{e^{-\lambda t}} = e^{-\lambda t'}.$$

Thus the probability of a radioactive atom's still being alive at time t' is independent of its age, a fact in contrast to every biological process.

*Articles marked with an asterisk are of a more difficult nature and may be omitted in a first reading.

Inserting $P(AB) = P(B)P(A|B)$ in (1), we find

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}. \tag{2}$$

Let us now consider ν events, $A_1, A_2, \cdots, A_\nu$, which exclude each other two-by-two and of which at least one occurs with certainty; in symbols $A_iA_k = O$, $i \neq k$, and $A_1 + A_2 + \cdots + A_\nu = E$. Furthermore, let X be an arbitrary event—not necessarily one of the events $A_1, A_2, \cdots, A_\nu$. It is then obvious that, together with the event X, one of the events $A_1, A_2, \cdots, A_\nu$, must occur; that is, if X has occurred, then either the event A_1X or the event A_2X or $\cdots$ or $A_\nu X$ has occurred; i.e., symbolically

$$X = A_1X + A_2X + \cdots + A_\nu X. \tag{3}$$

Since the events A_i and A_k excluded each other for $i \neq k$, the same applies to the events A_iX and A_kX. Thus from (3) we get

$$P(X) = P(A_1X) + P(A_2X) + \cdots + P(A_\nu X). \tag{4}$$

Inserting $P(A_iX) = P(A_i)P(X|A_i)$ in (4), we get

$$P(X) = P(A_1)P(X|A_1) + \cdots + P(A_\nu)P(X|A_\nu). \tag{5}$$

Choosing, in (2), $A = X$ and $B = A_i$, and inserting (5) for $P(X)$, we finally obtain

$$P(A_i|X) = \frac{P(A_i)P(X|A_i)}{P(X)} = \frac{P(A_i)P(X|A_i)}{P(A_1)P(X|A_1) + \cdots + P(A_\nu)P(X|A_\nu)}, \tag{6}$$

which is called **Bayes' theorem.**

To summarize: A certain event X has been observed, and we know that it has occurred as a consequence of one of ν events $A_1, \cdots, A_\nu$, which exclude each other two-by-two. The events $A_1, \cdots, A_\nu$ are called the "causes" or "hypotheses" of X, and (6) is said to give the probability of the event A_i after the event X has occurred.

Example 2. Let us apply Bayes' theorem to the example, § 1.5. The three hypotheses in this case are the three bureaus, $A = A_1$, $B = A_2$, $C = A_3$, and X is the event "at least one of the three drawers contains a gold coin." Thus we have

$$P(A_1) = P(A_2) = P(A_3) = 1/3.$$

Furthermore we have $P(X|A_1) = 1$, since both drawers in bureau A contain a gold coin; $P(X|A_2) = \frac{1}{2}$, since only one drawer in bureau B contains a gold coin; and, finally, $P(X|A_3) = 0$, since none of the

drawers in bureau C contains a gold coin. Inserting these numbers in (6) we get

$$P(B_1|X) = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3} \cdot 1 + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0} = \frac{1}{3}$$

in agreement with the result found in §1.5.

Exercise. Write down Bayes' theorem in the case $X = \sum_{i=1}^{m} A_i$, $m < \nu$.

*§3.9. **The multinomial law.** In Bernoulli's problem, §3.7, we considered the possibility that in each of ν observations there were only two possibilities, either the event A occurred or it did not occur. More generally, one also meets the problem that in each of ν observations we have more than two possibilities. Let us here consider the case in which in each observation, one and only one of k mutually exclusive events occurs: $A_1, A_2, \cdots, A_k$. Thus we assume

$$A_1 + A_2 + \cdots + A_k = E, \qquad A_i A_j = O \text{ for } i \neq j. \tag{1}$$

Let us denote by $\theta_i = P(A_i)$ the probability that A_i occurs. From (1) we then have

$$\theta_1 + \theta_2 + \cdots + \theta_k = 1. \tag{2}$$

We now ask for the probability, $P_{r_1, r_2, \cdots, r_k}$, that among the total $r_1 + r_2 + \cdots + r_k = \nu$ observations the event A_1 occurs r_1 times, A_2 r_2 times, $\cdots$, and A_k r_k times. This composite event may occur in a number of different ways corresponding to the number of different permutations of r_1 A_1's, r_2 A_2's, $\cdots$, and r_k A_k's. From (3.4.1) the probability of each one of these permutations is

$$\theta_1^{r_1} \theta_2^{r_2} \cdots \theta_k^{r_k}. \tag{3}$$

From (3.3.1) we then obtain the desired probability $P_{r_1, r_2, \cdots, r_k}$ by multiplying (3) by the number of ways in which we can select r_1 elements, r_2 elements, $\cdots$, and r_k elements out of a total number of $r_1 + r_2 + \cdots + r_k = \nu$ elements. This number is given by

$$\frac{\nu!}{r_1! r_2! \cdots r_k!}.$$

Consequently,

$$P_{r_1, r_2, \cdots, r_k} = \frac{\nu!}{r_1! r_2! \cdots r_k!} \theta_1^{r_1} \theta_2^{r_2} \cdots \theta^{r_k},$$

$$r_1 + r_2 + \cdots + r_k = \nu. \tag{4}$$

This formula is called the **multinomial law.** We see that it contains the binomial law as a special case for $k = 2$. By means of the multinomial formula and (2) we have in agreement with (3.5.3)

$$\sum_{r_1 + r_2 + \cdots + r_k = \nu} P_{r_1, r_2, \cdots, r_k} = (\theta_1 + \theta_2 + \cdots + \theta_k)^\nu = 1. \tag{5}$$

4.

RANDOM VARIABLES AND DISTRIBUTION FUNCTIONS

§ 4.1. In the following we shall deal exclusively with random phenomena in which the directly observed result is one or more numbers.[1] *A quantity which can be determined quantitatively and which in different observations of the same category can assume different values shall be called a random variable, symbolized by* x.[2] Thus the letter x not only symbolizes the quantity considered but also the way in which it has to be measured, the category (cf. § 1.6). Examples of random variables are the result of a throw of a die; the weight and height of one of a group of soldiers; and the range, elevation, and azimuth of a shot. In the last-mentioned example, x is not defined before the experimental arrangement has been specified, i.e., the type and position of the gun, the composition of the powder, the weight of the projectile, and so on. At first we shall assume that the random variables considered are **one dimensional,** i.e., that they can be characterized by one number.

§ 4.2. The probability that x assumes the value t we denote by $P(x = t)$, and the probability of x assuming a value in the interval $a < t \leqq b$ we denote by $P(a < x \leqq b)$ or simply $P(a, b)$. If we know $P(a, b)$ for all values of a and b, we say that we know the **distribution** of x or how x is **distributed.** We have a convenient way to describe a distribution in the so-called distribution function. By the **distribution function,**[3] $\Phi(t)$, *for a random variable* x we *understand that func-*

[1] Even though the directly observed result may not be a number, we can always symbolize a result by a number. For instance, in the example stated in § 1.3 we can associate the number 1 to the result "male" and the number 0 to the result "female."

[2] It is also called a **statistical** or **stochastic** variable or a **variate.** In French, **variable aléatoire;** in German, **zufällige Grösse.**

[3] $\Phi(t)$ is also called the **sum function** or the **cumulative, total,** or **integral distribution function,** in order to distinguish it from the function $\varphi(t)$ to be introduced later (cf. § 4.4). French, **fonction des probabilités totales;** German, **Verteilungsfunktion.** Often the distribution function is denoted by $F(t)$, but we shall use this for the corresponding empirical function (cf. Chapter 10), for the reasons discussed on p. 9.

tion which for all values of t is equal to the probability of $x \leqq t$:

$$\Phi(t) = P(-\infty < x \leqq t) = P(-\infty, t). \tag{1}$$

If at the same time we consider several random variables, $x, y, \cdots$, the corresponding distribution functions may be written $\Phi_x(t)$, $\Phi_y(t)$, $\cdots$. From (3.1.1) we have for $t_1 < t_2$

$$P(-\infty, t_2) = P(-\infty, t_1) + P(t_1, t_2),$$

and, inserting (1) into this, we have, due to $P(t_1, t_2) \geqq 0$

$$\Phi(t_1) \leqq \Phi(t_2). \tag{2}$$

Thus the distribution function $\Phi(t)$ is a never-decreasing function.

Example. A shot from a rifle has the probability of hitting of 1/2. We fire five shots. The random variable x is the number of hits. The probability of $P(x = r)$, $r = 0, 1, 2, 3, 4, 5$, is given by the binomial law (3.7.2), and thus

$$\Phi(t) = \begin{cases} 0 & \text{for} \quad -\infty < t < 0 \\ 1/32 & \text{for} \quad 0 \leqq t < 1 \\ 6/32 & \text{for} \quad 1 \leqq t < 2 \\ 16/32 & \text{for} \quad 2 \leqq t < 3 \\ 26/32 & \text{for} \quad 3 \leqq t < 4 \\ 31/32 & \text{for} \quad 4 \leqq t < 5 \\ 1 & \text{for} \quad 5 \leqq t < \infty. \end{cases}$$

The graph $u = \Phi(t)$ of this function is shown in Fig. 2. Random variables, such as the one mentioned here, which can assume only the values $r = 1, 2, \cdots, v$, with the probabilities P_r given by (3.7.2), are said to be **binomially distributed**.

We note in the last example that the distribution function $\Phi(t)$ is continuous for all values of t for which the probability of x assuming this value is 0 but that it is discontinuous in the other points, with jumps at these points equal to the probabilities of x assuming the corresponding values of t. Furthermore, we notice that at a discontinuity point $\Phi(t)$ is equal to its limit from the right. Finally, we note that $\Phi(t)$ has the limits 0 for $t \to -\infty$ and the limit 1 for $t \to \infty$.

Quite generally, it can be proved from axioms I to VI that every distribution function has these properties:[1]

$$\lim_{t \to \infty} \Phi(t) = \lim_{t \to \infty} P(-\infty < x \leqq t) = 1, \tag{3}$$

[1] See, e.g., Cramér, *Random Variables*, Chapter II.

i.e., $\Phi(t)$ approaches 1 without limit for $t \to \infty$.

$$\lim_{t \to -\infty} \Phi(t) = \lim_{t \to -\infty} P(-\infty < x \leqq t) = 0, \tag{4}$$

i.e., $\Phi(t)$ approaches 0 without limit for $t \to -\infty$.[1]

$$\lim_{t \to t_0} \Phi(t) = \Phi(t_0), \qquad (t > t_0), \tag{5}$$

i.e., $\Phi(t)$ approaches $\Phi(t_0)$ when t approaches t_0 from the *right-hand* side.

$$\lim_{t \to t_0} [\Phi(t_0) - \Phi(t)] = P(x = t_0), \qquad (t < t_0), \tag{6}$$

i.e., the difference $\Phi(t_0) - \Phi(t)$ approaches $P(x = t_0)$ when t approaches t_0 from the *left-hand* side. Thus, if t_0 is a continuity point for the function $\Phi(t)$, we have $P(x = t_0) = 0$.

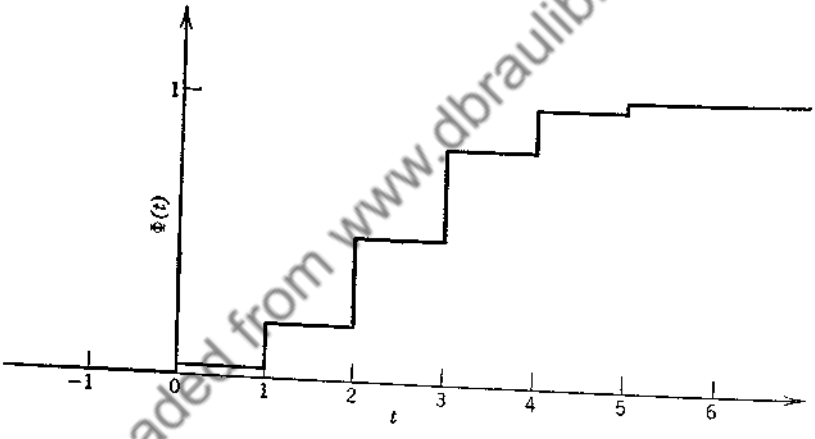

FIG. 2.

The distribution function $\Phi(t)$ of the random variables encountered in practice will, as a rule, belong to one of the two following types.

I. $\Phi(t)$ is piecewise constant (as in Fig. 2).

II. $\Phi(t)$ is continuous and piecewise differentiable with a continuous derivative.

In these cases we speak of discontinuous and continuous distribution functions respectively. In both cases the analytic expressions will

[1] It is taken for granted here that x can assume only finite values. If x can assume the values ∞ and $-\infty$ with a positive probability, we have $\lim_{t \to \infty} \Phi(t) < \Phi(\infty) = 1$ and $\lim_{t \to -\infty} \Phi(t) = \Phi(-\infty) > 0$. This can occur, e.g., when we go from one variable having a positive probability of assuming the value 0 to its logarithm which would then have a positive probability of assuming the value $-\infty$.

often contain one or more constants, called **parameters,** which we shall denote by Greek letters. The numerical values of these parameters are in each single case estimated from the observations, so that the theoretical distribution describes the observations in the best way (cf. Chapter 10).

§4.3. Discontinuous distribution functions. We say that a distribution function $\Phi(t)$ is **discontinuous** if it is piecewise constant, by which we mean that there exist certain points

$$\cdots < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < \cdots$$

at which $\Phi(t)$ makes positive jumps

$$\cdots, \quad \varphi_{-2}, \quad \varphi_{-1}, \quad \varphi_0, \quad \varphi_1, \quad \varphi_2, \quad \cdots$$

and that $\Phi(t)$ is constant between two consecutive values of t_i. The values t_i are often called the **spectrum** of the random variable x. If the spectrum consists of the values $0, 1, 2, \cdots$, it is more convenient to write simply i instead of t_i.

Thus the function $u = \Phi(t)$ is a step function lying between the lines $u = 0$, $u = 1$ and having steps of the heights φ_i at the jump points t_i. The number of steps can be either finite or infinite. From (4.2.6) the height of the jump, φ_i, is equal to the probability of x assuming the value t_i. For any other value of t we have $P(x = t) = 0$ due to the continuity of $\Phi(t)$ at this point.

Now we have

$$\sum_{i=q}^{p} \varphi_i = \Phi(t_p) - \Phi(t_{q-1}).$$

Letting $q \to -\infty$ here we obtain from (4.2.4)

$$\Phi(t_p) = \sum_{i=-\infty}^{p} \varphi_i, \tag{1}$$

and next letting $p \to \infty$ here we obtain from (4.2.3)

$$\sum_{i=-\infty}^{\infty} \varphi_i = 1. \tag{2}$$

Example 1. The simplest example of a discontinuous distribution is that in which the random variable x assumes with certainty one definite value only; let us call it μ. Thus the spectrum consists of

only this single value, and we have

$$t_0 = \mu$$
$$\varphi_0 = P(x = \mu) = 1 \tag{3}$$
$$\Phi(t) = P(x \leqq t) = \epsilon(t - \mu) = \begin{cases} 0 & \text{for} \quad t < \mu \\ 1 & \text{for} \quad t \geqq \mu. \end{cases}$$

Here we have one parameter, μ. This distribution is obviously precisely that which corresponds to a causal description (cf. p. 3). Therefore we shall call it the **causal distribution.**

Example 2. An example of discontinuous distributions often met in practice, especially in biology, is the binomial distribution which we have considered already in the example, § 4.2. In the general **binomial distribution** we have from (3.7.2) and (3.7.4)

$$i = 0, 1, 2, \cdots, \nu$$
$$\varphi_i = P_i = \binom{\nu}{i} \theta^i (1 - \theta)^{\nu - i} \tag{4}$$
$$\sum_{i=0}^{\nu} \varphi_i = 1.$$

Here θ and ν are the parameters, but in most cases only θ is unknown, ν being given by the problem itself.

All random variables denoting a number are also examples of discontinuous distributions. In the following we shall give three such distributions which are important in many practical applications.

Example 3. In **Poisson's distribution** the spectrum consists of all non-negative integers, and the distribution is given by

$$i = 0, 1, 2, \cdots$$
$$\varphi_i = e^{-\mu} \frac{\mu^i}{i!}, \qquad \mu > 0 \tag{5}$$
$$\sum_{i=0}^{\infty} \varphi_i = e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^i}{i!} = e^{-\mu} e^{\mu} \equiv 1,$$

where μ is a parameter.

Poisson's distribution is often met, as shown by the following examples.

Example 4. Let the random variable x be the number of calls during a time interval t at a telephone switchboard. If we assume

(*a*) that the probability of a call in the time dt is proportional to dt, $\lambda\, dt$, and (*b*) that the calls are independent, x is Poisson distributed.

Denoting the probability of i calls in the time t by $P_i(t)$ we have that $P_i(t + dt)$ is equal to the product of the probabilities of i calls in t and 0 in dt plus the product of the probabilities of $i - 1$ calls in t and 1 in dt. (dt being a differential, the probability of more than one call in dt is of higher order and therefore negligible.) Therefore, from our assumptions, we have

$$P_i(t + dt) = P_i(t)(1 - \lambda\, dt) + P_{i-1}(t)\lambda\, dt,$$

from which

$$\frac{dP_i(t)}{dt} = \lambda(P_{i-1}(t) - P_i(t)). \tag{6}$$

It is easily seen that the solution of this infinite system of differential equations is uniquely given by Poisson's formula, with $\mu = \lambda t$,

$$P_i(t) = e^{-\lambda t}\frac{(\lambda t)^i}{i!}, \quad P_i(0) = \begin{cases} 1 & i = 0 \\ 0 & i > 0. \end{cases} \tag{7}$$

(Putting $P_i(t) = \dfrac{(\lambda t)^i}{i!} f(t)$, (6) shows that $f(t)$ is uniquely given by $e^{-\lambda t}$ for the initial condition given in (7).) The formula (7) is of great importance, e.g., in designing telephone switchboards.[1]

Example 5. Poisson's distribution given in (5) has many other applications since the assumptions (*a*) and (*b*) can be carried over to other fields. Thus (5) also gives the probability of finding i radioactive atoms decaying in the time interval t; the probability that i cosmic-ray particles trip a Geiger-Müller counter in the time interval t; the probability of a store's selling i pieces of a certain type of merchandise in the time interval t; the probability of i cars passing a certain street in the time interval t; the probability of i suicides occurring in the time interval t; the probability of finding i corpuscles within the field of area t of a microscope, and so forth.

Example 6. Pascal's distribution is given by

$$i = 0, 1, 2, \cdots$$

$$\varphi_i = \frac{1}{1 + \mu}\left(\frac{\mu}{1 + \mu}\right)^i, \qquad \mu > 0 \tag{8}$$

$$\sum_{i=0}^{\infty} \varphi_i = \frac{1}{1 + \mu}\sum_{i=0}^{\infty}\left(\frac{\mu}{1 + \mu}\right)^i = \frac{1}{1 + \mu}\cdot\frac{1}{1 - \dfrac{\mu}{1 + \mu}} \equiv 1,$$

[1] See, e.g., Fry, *Probability and Its Engineering Uses.*

where μ is a parameter. (In the theory of cosmic rays this is also called Furry's distribution.)

Example 7. Pólya's distribution is defined as follows, μ and β being parameters,

$$i = 0, 1, 2, \cdots,$$

$$\varphi_0 = (1 + \beta\mu)^{-1/\beta}, \qquad \beta > 0,\ \mu > 0$$

$$\varphi_i = \left(\frac{\mu}{1 + \beta\mu}\right)^i \frac{1(1 + \beta) \cdots (1 + (i - 1)\beta)}{i!}\varphi_0, \qquad i \geqq 1$$

$$\sum_{i=0}^{\infty} \varphi_i = \varphi_0 \sum_{i=0}^{\infty} \binom{-1/\beta}{i} \left(\frac{-\beta\mu}{1 + \beta\mu}\right)^i \tag{9}$$

$$= (1 + \beta\mu)^{-1/\beta} \left(1 - \frac{\beta\mu}{1 + \beta\mu}\right)^{-1/\beta} \equiv 1.$$

Here we have used the general binomial theorem and the fact that

$$\frac{1(1 + \beta) \cdots (1 + (i - 1)\beta)}{i!} = (-\beta)^i \binom{-1/\beta}{i}.$$

(Because of this relation Pólya's distribution is also called the **negative binomial distribution.**)

One should notice that the Pólya distribution contains two parameters, μ and β, in contrast to both the Poisson and the Pascal distributions which have only one, μ. It is therefore more flexible than these two distributions, which are, incidentally, only special cases of the Pólya distribution. In fact, from (9), we find that passing to the limit $\beta \to 0$ we obtain just the Poisson distribution

$$\varphi_i \underset{\beta \to 0}{\longrightarrow} \begin{cases} e^{-\mu} & \text{for} \quad i = 0 \\ e^{-\mu} \dfrac{\mu^i}{i!} & \text{for} \quad i \geqq 1. \end{cases} \tag{10}$$

For $\beta = 1$ we obtain just Pascal's distribution

$$\varphi_i \underset{\beta=1}{=} \frac{1}{1 + \mu} \left(\frac{\mu}{1 + \mu}\right)^i. \tag{11}$$

§4.4. Continuous distributions. Strictly speaking all random variables are discontinuous because in practice we can only measure multiples of the smallest unit that the measuring apparatus can measure. However, if this unit is very small compared with the variations

of x observed, we may to a good approximation abstract from this fact and treat x as a continuous variable.

We call a distribution **continuous** if $\Phi(t)$ is continuous everywhere and is piecewise differentiable with a continuous derivative, $\Phi'(t) = \varphi(t)$ (i.e., $\varphi(t)$ exists and is continuous except possibly at certain points which occur at most a finite number of times in any finite interval). Thus we have

$$\int_{t_1}^{t_2} \varphi(t)\, dt = \Phi(t_2) - \Phi(t_1) = P(t_1 < x \leqq t_2); \tag{1}$$

i.e., the probability of x assuming a value in a certain interval is given by the area between the curve $u = \varphi(t)$ and the given interval on the t-axis. The function $\varphi(t)$ is called the **probability density** of x or its **differential distribution function.**[1] All values of t for which $\varphi(t) \neq 0$ are said to make up the **spectrum** of x. Since $\varphi(t)$ is the derivative of a never-decreasing function, $\varphi(t) \geqq 0$ for all values of t. If φ has only one maximum, the distribution is called **unimodal,** if two, **bimodal,** and so forth. In (1) letting $t_1 \to -\infty$ and putting $t_2 = t$ we obtain from (4.2.4)

$$\Phi(t) = \int_{-\infty}^{t} \varphi(t)\, dt, \tag{2}$$

which is analogous to (4.3.1). In (2) letting $t \to \infty$ we obtain from (4.2.3)

$$\int_{-\infty}^{\infty} \varphi(t)\, dt = 1, \tag{3}$$

which is analogous to (4.3.2).

In (1) putting $t_1 = t$ and $t_2 = t + \Delta t$ we have

$$P(t, t + \Delta t) = \int_{t}^{t+\Delta t} \varphi(t)\, dt,$$

which by means of the theorem of mean values gives

$$P(t, t + \Delta t) = \varphi(\xi)\, \Delta t, \tag{4}$$

in which ξ lies between t and $t + \Delta t$.

Because of (4) we can say that $\varphi(t)\, dt$ gives the probability of x assuming a value in the "infinitely small" interval t to $t + dt$. We wish to stress that it is $\varphi(t)\, dt$ and not $\varphi(t)$ itself that gives the proba-

[1] $\varphi(t)$ is also called the **frequency function.** This expression is, however, somewhat misleading because $\varphi(t)$ refers to the theoretical model, not to the observed frequencies. French, **densité de probabilité;** German, **Wahrscheinlichkeitsdichte.** Unfortunately, the terminology is not yet fixed, so that one has to be careful to see whether "distribution function" applies to the total or to the differential distribution function.

bility. Therefore one often gives the probability density in the form

$$d\Phi = \varphi(t)\, dt, \tag{5}$$

which is called the **probability differential.** Furthermore, we stress that in a continuous distribution we ask for the probability of x assuming a value *between* t and $t + dt$ and not for the probability of x assuming *exactly* the value of t. We shall now give some examples of continuous distributions.

Example 1. One of the simplest continuous distributions is the **rectangular distribution** given by

$$d\Phi(t) = \varphi(t)\, dt = \begin{cases} 0 & \text{for} \quad t < \alpha \\ \dfrac{1}{\beta - \alpha}\, dt & \text{for} \quad \alpha < t < \beta, \\ 0 & \text{for} \quad \beta < t \end{cases} \tag{6}$$

in which α and β are two parameters. Since $\varphi(t)$ is constant, x is also said to be *uniformly distributed* within the interval (α, β). Check that $\int_{-\infty}^{\infty} \varphi(t)\, dt = 1$, and find $\Phi(t)$. Draw the graphs of $\varphi(t)$ and $\Phi(t)$. Show that keeping α fixed and letting $\beta \to \alpha$, $\Phi(t) \to \epsilon(t - \alpha)$, in which $\epsilon(t - \alpha)$ is the causal distribution given in Example 1, § 4.3, with $\alpha = \mu$.

Example 2. Let the random variable x be the life time of a radioactive atom. Since the lifetime is a non-negative number, we have

$$P(x \leqq t) = \Phi(t) = 0 \qquad \text{for} \qquad -\infty < t < 0.$$

If $x \leqq t$ (for $t \geqq 0$) this means that the atom must have decayed at a previous time. Since the complementary event is the event that the atom is still "alive" at the time t we obtain from (3.5.4) and (3.7.5)

$$P(x \leqq t) = \Phi(t) = 1 - e^{-\lambda t} \qquad \text{for} \qquad 0 \leqq t < \infty,$$

i. e.,

$$d\Phi(t) = \varphi(t)\, dt = \begin{cases} 0 \cdot dt & \text{for} \quad t \leqq 0 \\ e^{-\lambda t} \lambda\, dt & \text{for} \quad 0 \leqq t < \infty. \end{cases} \tag{7}$$

Check that $\int_{-\infty}^{\infty} \varphi(t)\, dt = 1$, and draw the graphs of $\varphi(t)$ and $\Phi(t)$.

We could also have obtained the result (7) directly, since it gives the probability of the atom's decaying in the time between t and $t + dt$. However, this probability is the product of the probability of the atom's still being alive at the time t, which from (3.7.5) is $e^{-\lambda t}$, and the probability of the atom's decaying in dt, which, due to our assumption,

is $\lambda\, dt$ (cf. Example 4, §4.3), which shows that (7) also gives the distribution of the time between two consecutive calls.

Example 3. The most important distribution in practice is the **normal distribution** given by

$$d\Phi(t) = \varphi(t)\, dt = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt \qquad (-\infty < t < \infty), \tag{8}$$

in which μ and σ are two parameters. The factor $1/\sqrt{2\pi}\,\sigma$ is called the **normalization factor,** i.e., that factor which makes $\int_{-\infty}^{\infty} \varphi(t)\, dt \equiv 1$. (We shall prove this in the detailed discussion in Chapter 7.) Draw the graphs for $\varphi(t)$ and $\Phi(t)$. Next show that, by letting $\sigma \to 0$, $\Phi(t) \to \epsilon(t-\mu)$ as given in Example 1, §4.3.

Example 4. Cauchy's distribution is given by

$$d\Phi(t) = \varphi(t)\, dt = \frac{1}{\pi\alpha} \frac{1}{1 + \dfrac{(t-\mu)^2}{\alpha^2}}\, dt \qquad (-\infty < t < \infty), \tag{9}$$

in which μ and α are two parameters. Check that $\int_{-\infty}^{\infty} \varphi(t)\, dt \equiv 1$, and draw the graphs of $\varphi(t)$ and $\Phi(t)$. Show that, for $\alpha \to 0$, $\Phi(t) \to \epsilon(t-\mu)$, as before. This distribution is found, e.g., in the intensity distribution of spectral lines and in the theory of scattering and capture processes of atomic nuclei.

Example 5. Laplace's distribution is given by

$$d\Phi(t) = \varphi(t)\, dt = \frac{1}{2\alpha} e^{-\frac{|t-\mu|}{\alpha}}\, dt \qquad (-\infty < t < \infty), \tag{10}$$

in which μ and α are two parameters. Check that $\int_{-\infty}^{\infty} \varphi(t)\, dt \equiv 1$, and draw the graphs of $\varphi(t)$ and $\Phi(t)$. Show that, by letting $\alpha \to 0$, $\Phi(t) \to \epsilon(t-\mu)$ also.

Exercise 1. Let $x_1, x_2, \cdots, x_\nu$ be ν independent random variables with the same distribution function $\Phi(t)$. The largest number among the x's is again a random variable. Show that its distribution function, $\Phi_{\max}(t)$, is given by

$$d\Phi_{\max}(t) = \nu\Phi^{\nu-1}(t)\varphi(t)\, dt = d\Phi^{\nu}(t). \tag{11}$$

***Example 6.** Since it is inconvenient in proving general theorems to repeat analogous proofs by treating discontinuous and continuous distributions separately, it is often useful to regard the continuous distributions as the general case containing the discontinuous distributions as special limiting cases. As shown in the previous examples, the simplest discontinuous distribution, the causal one, $\Phi(t) =$

$\epsilon(t-\mu)$, may be obtained in a variety of ways as a limit of a continuous distribution. Dirac[1] has therefore introduced a fictitious function giving the "probability density," $\delta(t-\mu)$, of $\epsilon(t-\mu)$, a function which obviously is 0 for all $t \neq \mu$ and ∞ for $t = \mu$ in such a way that the whole integral is 1:

$$\frac{\text{“}d\epsilon(t-\mu)\text{”}}{dt} = \delta(t-\mu) = \delta(\mu - t)$$

$$= \begin{cases} 0 & \text{for} \quad t \neq \mu \\ \infty & \text{for} \quad t = \mu \end{cases} \tag{12}$$

$$\int_{-\infty}^{\infty} \delta(t-\mu)\, dt = \int_{\mu-\epsilon_1}^{\mu+\epsilon_2} \delta(t-\mu)\, dt = 1,$$

in which ϵ_1 and ϵ_2 are arbitrary positive numbers. Of course, such a "function" does not exist mathematically, and its introduction is therefore a non-rigorous procedure. However, first, it is only a shorthand expression for the result of a limiting process, and the final results obtained are always correct; and, second, the method may be made rigorous by means of a more complicated concept, the Stieltjes' integral (cf. § 4.7). Consequently, the δ-function of Dirac, being much more "anschaulich" and being more convenient in application than either the strict limiting processes or the rigorous, but less known, Stieltjes' integral, has been widely adopted, especially in quantum theory, where "vigor" of a mathematical tool is more important than its "rigor." However, it is obvious that its application demands a certain tact in order to lead to the correct final results.

*Exercise 2. Show that any discontinuous distribution function may be written

$$\Phi(t) = \sum_{i=-\infty}^{\infty} \varphi_i \epsilon(t-\mu_i) \tag{13}$$

Next show that by means of the Dirac δ-function this may be written as a *continuous* distribution with *probability density:*

$$d\Phi(t) = \varphi(t)\, dt = \sum_{i=-\infty}^{\infty} \varphi_i\, \delta(t-\mu_i)\, dt. \tag{14}$$

*Exercise 3. Prove the following properties of the δ-function:

$$\int_{-\infty}^{\infty} f(t)\delta(t-\mu)\, dt = f(\mu) \tag{15}$$

for any continuous function $f(t)$;

$$\int_{-\infty}^{\infty} \delta(t-\mu_1)\delta(t-\mu_2)\, dt = \delta(\mu_1 - \mu_2) \tag{16}$$

and, by means of partial integration,

$$\int_{-\infty}^{\infty} f(t)\delta'(t-\mu)\, dt = -f'(\mu). \tag{17}$$

§ 4.5. The probability density $\varphi(t)$ is, when it exists, a handier as well as a simpler means for the study of a random variable than the

[1] P. A. M. Dirac, *Principles of Quantum Mechanics*, First edition, London, 1930.

distribution function $\Phi(t)$. However, it must be remembered that the distribution function is the general concept which can always be applied.

Most of the random variables found in practice will have purely discontinuous or purely continuous distribution functions. But there also occur in practice random variables which have mixed distribution functions, i.e., which have both discontinuity points and continuity intervals. Thus in atomic theory the energy of an atom is a random variable and its spectrum often has both a discontinuous and a continuous part. As another example we may mention the following. A 40-year old person purchases a so-called 25-payment life policy, by which it is understood that a certain amount of money is paid to him when he reaches 65 years or to his beneficiary immediately after his death. The time interval between the day on which the policy becomes effective and the day of payment is obviously a random variable with a distribution function equal to zero for $t < 0$, increasing continuously from $t = 0$ to $t = 25$ and jumping discontinuously at $t = 25$ to the value 1.

Finally it must be mentioned that theoretically we have still another type of distribution function in which $\Phi(t)$ is continuous everywhere, is differentiable "almost" everywhere with $\Phi'(t) \equiv 0$, but nevertheless increases from 0 to 1.[1] Such "pathological" distributions are, of course, never found in practice.

§ 4.6. In a convenient and often applied mechanical picture we can say that a distribution function defines a certain mass distribution on the t-axis. Thus we think of this axis as being an infinitely thin rod, the various parts of which are coated with mass of varying density. In addition, at certain isolated points, mass particles are concentrated. Putting the total amount of mass = 1 and denoting by $\Phi(t)$ the mass lying to the left of and including the point t we see that $\Phi(t)$ has exactly the same properties as a distribution function. Thus the discontinuity points of $\Phi(t)$ are the points at which mass particles are situated. In the continuity points the mass density is given by $\Phi'(t) = \varphi(t)$. It is from this mechanical analogy that the expression *probability density* originates. Often one also speaks directly of *probability mass*, meaning the magnitude of probability.

*§ 4.7. As already mentioned it is convenient to treat discontinuous and continuous distributions in the same manner. The method of Dirac indicated in § 4.4, being convenient but not rigorous, may be made rigorous by means of the so-called Stieltjes' integral, which is a simple generalization of the usual integral.[2]

[1] See, e.g., Titchmarsh, *Theory of Functions*, § 11.72, London, 1932.

[2] For a more detailed discussion we refer, e.g., to Cramér, *Mathematical Methods*

Let $F(t)$ be a never-decreasing real function in the interval $a \leq t \leq b$, and let $g(t)$ be real and continuous in the same interval. We divide the interval from a to b into n parts by means of the division points

$$a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$$

and choose in each of these subintervals an arbitrary point

$$\xi_\nu, \qquad t_\nu \leqq \xi_\nu \leqq t_{\nu+1}.$$

Then it may be shown, exactly as is shown by ordinary integrals, that for a sufficiently fine division the sum

$$\sum_{\nu=0}^{n-1} g(\xi_\nu)(F(t_{\nu+1}) - F(t_\nu))$$

will deviate by an arbitrarily small amount from a certain limit. This limit is called the Stieltjes' integral and is denoted by

$$\int_a^b g(t)\, dF(t). \tag{1}$$

Exercise 1. Show that, if $F(t)$ has a continuous derivative, then the Stieltjes' integral reduces to the ordinary integral

$$\int_a^b g(t)\, dF(t) = \int_a^b g(t)F'(t)\, dt. \tag{2}$$

Thus, especially if $F(t) = t$, (1) reduces to the ordinary integral of $g(t)$.

Exercise 2. Show that, if $F(t)$ is a step-function with the step-points t_i and steps F_i (cf. §4.3), then the Stieltjes' integral reduces to the sum

$$\int_a^b g(t)\, dF(t) = \sum_i g(t_i)F_i. \tag{3}$$

It may easily be shown that all elementary rules for definite integrals also hold for Stieltjes' integrals. Next the integral

$$\int_{-\infty}^{\infty} g(t)\, dF(t) \tag{4}$$

is defined in the usual way. If in (4) we put $g(t) = 1$ and $F(t) = \Phi(t)$, it follows from (4.2.3) and (4.2.4) that

$$\int_{-\infty}^{\infty} dF(t) = 1. \tag{5}$$

This single formula obviously contains the separate formulae (4.3.2) and (4.4.3).

of Statistics, § 7.5; or D. V. Widder, *The Laplace Transform*, Chapter 1, Princeton, 1946.

However, although the Stieltjes' integral is the ideal tool in probability, we shall not apply it here since most students of applied mathematics are not yet familiar with it.

*§4.8. By a function $y = f(x)$ we understand a new random variable, y, assuming the value $f(t)$, when x assumes the value t. If the condition $\frac{df}{dt} > 0$, or $\frac{df}{dt} < 0$, is satisfied for all t we have

$$\Phi_x(t) = P(x \leqq t) = \begin{cases} P(y = f(x) \leqq f(t) = u) \\ \quad = \Phi_y(u) & \text{for } \frac{df}{dt} > 0 \\ P(y = f(x) \geqq f(t) = u) \\ \quad = 1 - \Phi_y(u) & \text{for } \frac{df}{dt} < 0. \end{cases}$$

Differentiating on both sides we find in both cases that

$$d\Phi_x = \varphi_x(t)\,dt = \varphi_x(t(u))\left|\frac{dt}{du}\right| du$$
$$= \varphi_y(u)\,du = d\Phi_y. \tag{1}$$

This formula plays a very important role in many practical applications of probability.

Example. Let the random variable be the numerical value, v, of the velocity of a molecule with mass m in a gas at absolute temperature T. The distribution of v is given by the **Maxwell-Boltzmann law:**

$$d\Phi_v = \varphi_v(t)\,dt = \alpha t^2 e^{-\beta t^2}\,dt, \quad \beta = \frac{m}{2kT} \quad (0 \leqq t < \infty), \tag{2}$$

in which k is the physical constant called *Boltzmann's constant* and α is the normalization factor. (Cf. Exercise 3, §7.4.)

The distribution of the kinetic energy, $E = \frac{1}{2}mv^2 = \gamma v^2$, is then, from (1), given by

$$d\Phi_E = \varphi_E(u)\,du = \frac{\alpha}{\gamma} u e^{-(\beta/\gamma)u} \frac{1}{2\sqrt{\gamma u}}\,du = \alpha' u^{\frac{1}{2}} e^{-\beta' u}\,du$$
$$(0 \leqq u < \infty), \tag{3}$$

where $\alpha' = \alpha/2\gamma^{3/2}$ and $\beta' = \beta/\gamma = 1/kT$ are two new constants.

Exercise. Show by means of Appendix 1 that

$$\alpha = \frac{4}{\sqrt{\pi}}\beta^{3/2} = \sqrt{\frac{2}{\pi}}\left(\frac{m}{kT}\right)^{3/2}.$$

and thus

$$\alpha' = \frac{2}{\sqrt{\pi}}\left(\frac{1}{kT}\right)^{3/2}.$$

§4.9. We shall now extend our formulae to two- and many-dimensional random variables; i.e., random variables whose specifications are given by two or more numbers. Since all concepts and formulae may immediately be generalized from two to ν dimensions we shall state them for only two dimensions, denoting the random variable $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$.

Example. If in tossing a coin we denote heads by 1 and tails by 0, and if we play with 2 coins, $\boldsymbol{z}$ can assume the values (1, 1), (0, 1), (1, 0), and (0, 0).

If, as in this example, we are investigating the distributions of 2 one-dimensional variables it is an obvious idea to interpret them as components of 1 two-dimensional variable. We can then interpret $\boldsymbol{x}$ and $\boldsymbol{y}$ as cartesian coordinates in a plane and ask for the probability of $(\boldsymbol{x}, \boldsymbol{y})$ lying within a given region in this plane. The mechanical picture from §4.6 may be generalized to this case immediately assuming a mass of total amount 1 spread out continuously over the plane or concentrated in certain points or lines.

By the **joint distribution function** $\Phi(t, u)$, or simply the distribution function, of a two-dimensional random variable $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ we understand that function which for all values of t and u is equal to the probability of $\boldsymbol{x} \leqq t$ and $\boldsymbol{y} \leqq u$:

$$\Phi(t, u) = P(-\infty < \boldsymbol{x} \leqq t, -\infty < \boldsymbol{y} \leqq u). \tag{1}$$

As in the one-dimensional case it may be shown from our axioms I to VI that the following natural generalizations of (4.2.3) and (4.2.4) hold true.[1]

$$\lim_{\substack{t \to \infty \\ u \to \infty}} \Phi(t, u) = 1 \tag{2}$$

$$\lim_{t \to -\infty} \Phi(t, u_0) = 0 \tag{3}$$

$$\lim_{u \to -\infty} \Phi(t_0, u) = 0 \tag{4}$$

for any fixed t_0 and u_0.

Exercise 1. Show that for $a > 0$ and $b > 0$ we have $\Phi(t, u + b) \geqq \Phi(t, u)$; $\Phi(t + a, u) \geqq \Phi(t, u)$; $\Phi(t + a, u + b) \geqq \Phi(t, u + b)$; $\Phi(t + a, u + b) \geqq \Phi(t + a,$

[1] Here it is also assumed that $\boldsymbol{x}$ and $\boldsymbol{y}$ can assume only finite values (cf. footnote 1, p. 28).

$u)$; $\Phi(t + a, u + b) \geqq \Phi(t, u)$, and $\Phi(t + a, u + b) - \Phi(t + a, u) \geqq \Phi(t, u + b) - \Phi(t, u)$. Next show that $P(a_1 < x \leqq a_2, b_1 < y \leqq b_2) = \Phi(a_2, b_2) + \Phi(a_1, b_1) - \Phi(a_1, b_2) - \Phi(a_2, b_1)$.

Exercise 2. Prove that

$$P(x + y > a + b) \leqq P(x > a) + P(y > b)$$

and

$$P(|x + y| > a + b) \leqq P(|x| > a) + P(|y| > b).$$

Corresponding to our treatment of one-dimensional variables we shall also treat here two types of distributions, the discontinuous and the continuous, separately.

§4.10. Let two series of numbers be given:

$$\cdots < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < \cdots$$

$$\cdots < u_{-2} < u_{-1} < u_0 < u_1 < u_2 < \cdots.$$

We say that the two-dimensional random variable z has a discontinuous distribution function if the possible values of z are all of the form (t_i, u_j). Let the corresponding probabilities be φ_{ij}. In every rectangle $(t_p \leqq t < t_{p+1}, u_q \leqq u < u_{q+1})$ the distribution function of z will be constant and equal to

$$\Phi(t_p, u_q) = \sum_{i=-\infty}^{p} \sum_{j=-\infty}^{q} \varphi_{ij}, \tag{1}$$

from which it follows by means of (4.9.2) that

$$\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \varphi_{ij} = 1. \tag{2}$$

Example. Let us assume that z can have only the values (0, 0), (2, 1), (1, 2), and (3, 3) with the respective probabilities, 1/4, 1/8, 3/8, and 1/4. The graphical picture and values of the distribution function are shown on Figs. 3 and 4.

Exercise. Find the distribution function of the variable in the example, §4.9.

§4.11. The probability that x assumes the value t_i independently of which value y assumes is symbolized by $\varphi_{i\cdot}$ and is given by

$$\varphi_{i\cdot} = \sum_{j=-\infty}^{\infty} \varphi_{ij}. \tag{1}$$

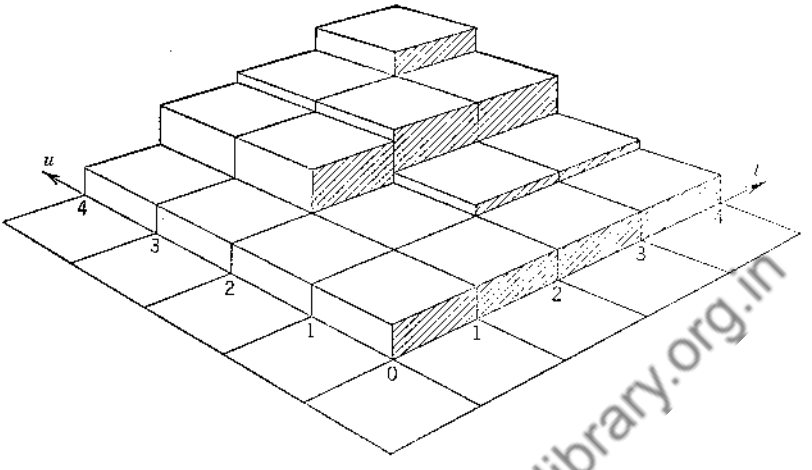

FIG. 3.

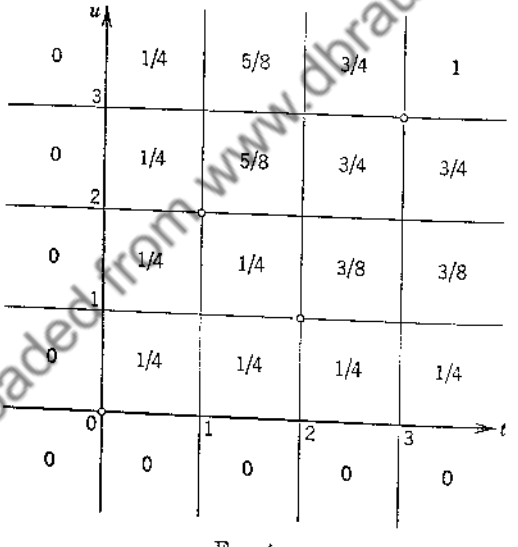

FIG. 4.

Analogously, the probability of y assuming the value u_j independently of which value x assumes is symbolized by $\varphi_{\cdot j}$ and is given by

$$\varphi_{\cdot j} = \sum_{i=-\infty}^{\infty} \varphi_{ij}. \tag{2}$$

The distributions $\varphi_{i\cdot}$ and $\varphi_{\cdot j}$ are called the **marginal distributions** of x and y respectively.

The conditioned probability of x assuming the value t_i under the

assumption that y has assumed the value u_j is symbolized by $\varphi_{i|j}$. Analogously, $\varphi_{j\,i}$ denotes the conditioned probability of y assuming the value u_j under the assumption that x has assumed the value t_i. From the multiplication law, V, we then have

$$\varphi_{ij} = \varphi_{i.}\varphi_{j|i} = \varphi_{.j}\varphi_{i|j}. \tag{3}$$

Exercise 1. Show that

$$\sum_{i=-\infty}^{\infty} \varphi_{i|j} = \sum_{j=-\infty}^{\infty} \varphi_{j|i} = 1. \tag{4}$$

If, for all values of i and j, $\varphi_{i|j} = \varphi_{i.}$, which from (3) implies $\varphi_{j|i} = \varphi_{.j}$, we say that x and y are **stochastically independent** or shortly **independent** (cf. p. 20), and (3) then reduces to

$$\varphi_{ij} = \varphi_{i.}\varphi_{.j}. \tag{5}$$

Exercise 2. Conversely, show that, if (5) holds true, x and y are independent.

Exercise 3. If x and y are independent, show that $z = x + y$ has a distribution given by

$$P(z = x + y = v_k) = \sum_{t_i + u_j = v_k} \varphi_{i.}\varphi_{.j}. \tag{6}$$

Exercise 4. Show by means of (6) that, if x and y are binomially distributed with the parameters ν_1, θ and ν_2, θ, respectively (cf. Example 2, § 4.3), then $z = x + y$ is also binomially distributed with the parameters $\nu = \nu_1 + \nu_2$, θ.

Exercise 5. Show by means of (6) that, if x and y both have the Poisson distributions with parameters μ_1 and μ_2, respectively, then $z = x + y$ also has the Poisson distribution with the parameter $\mu = \mu_1 + \mu_2$.

§ 4.12. We say that the two-dimensional random variable z has a continuous distribution function if there exists a piecewise continuous function $\varphi(t, u) \geqq 0$ (i.e., φ exists and is continuous apart from possibly certain points on a finite number of curves) so that the probability of z assuming a value in the region ω of the tu-plane is given by the plane integral

$$P(z \text{ in } \omega) = \int\int_{\omega} \varphi(t, u)\, dt\, du. \tag{1}$$

$\varphi(t, u)$ is called the **probability density** (or the **correlation function**) of x and y. In analysis it is shown that

$$\varphi(t, u) = \frac{\partial^2 \Phi(t, u)}{\partial t\, \partial u} \tag{2}$$

at the points at which $\varphi(t, u)$ exists and is continuous.

Thus the probability of z assuming a value in a certain region is given by the volume lying between the surface, $v = \varphi(t, u)$, and the

given area in the tu-plane (cf. the one-dimensional case). We say that $\varphi(t, u)\, dt\, du$ gives the probability of $\boldsymbol{z}$ assuming a value in the "infinitely small" interval $t \leqq \boldsymbol{x} \leqq t + dt$, $u \leqq \boldsymbol{y} \leqq u + du$. We wish to stress that it is $\varphi(t, u)\, dt\, du$ and not $\varphi(t, u)$ itself which gives the probability. In analogy with (4.4.5) the distribution is often given in the form

$$d\Phi = \varphi(t, u)\, dt\, du, \tag{3}$$

which is also called the **probability differential.**

From (1), or (2), it follows that

$$\Phi(t, u) = \int_{-\infty}^{t} \int_{-\infty}^{u} \varphi(t, u)\, dt\, du, \tag{4}$$

and from (4.9.2)

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(t, u)\, dt\, du = 1. \tag{5}$$

Example 1. Let us consider the firing of a cannon, and let us assume that the azimuthal deviation $\boldsymbol{x}$ and the height deviation $\boldsymbol{y}$ from a target having x, y coordinates μ_x, μ_y are independent and normally distributed with parameters μ_x, σ_x and μ_y, σ_y respectively (cf. Example 3, § 4.4). Then from the multiplication law the distribution of $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ is given by

$$d\Phi = \varphi(t, u)\, dt\, du = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\left\{ \frac{(t-\mu_x)^2}{2\sigma_x{}^2} + \frac{(u-\mu_y)^2}{2\sigma_y{}^2} \right\} \right] dt\, du$$
$$(-\infty < t < \infty,\ -\infty < u < \infty). \tag{6}$$

The general two-dimensional normal distribution can be obtained from (6) by adding a cross term in the exponential and changing the normalization factor correspondingly:

$$d\Phi = \varphi(t, u)\, dt\, du = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$
$$\exp\left[-\frac{1}{1-\rho^2} \left\{ \frac{(t-\mu_x)^2}{2\sigma_x{}^2} - \rho\, \frac{(t-\mu_x)(u-\mu_y)}{\sigma_x\sigma_y} + \frac{(u-\mu_y)^2}{2\sigma_y{}^2} \right\} \right] dt\, du$$
$$(-\infty < t < \infty,\ -\infty < u < \infty). \tag{7}$$

Here ρ is a new parameter for which $-1 \leqq \rho \leqq 1$. For $\rho = \pm 1$, $\boldsymbol{x}$ and $\boldsymbol{y}$ become proportional since then the total probability mass is concentrated on the straight line $u = \mu_y \pm \frac{\sigma_y}{\sigma_x}(t - \mu_x)$. (Later, in Chapter 7, we shall discuss this distribution in greater detail.)

Example 2. Another continuous distribution is given by

$$d\Phi = \varphi(t, u)\, dt\, du$$
$$= \frac{1}{\pi} \frac{1}{(1 + t^2 + u^2)^2}\, dt\, du. \tag{8}$$

Check that (5) is fulfilled in this case.

§ 4.13. The probability of x assuming a value between $-\infty$ and t independently of which value y assumes is given by

$$\Phi_x(t) = \int_{-\infty}^{t} dt \int_{-\infty}^{\infty} \varphi(t, u)\, du. \tag{1}$$

$\Phi_x(t)$, or briefly $\Phi(t)$, is called the **marginal distribution** of x. In analysis it is shown that, if we assume that the function

$$\varphi_x(t) = \int_{-\infty}^{\infty} \varphi(t, u)\, du, \tag{2}$$

or briefly $\varphi(t)$, has at most a finite number of discontinuities, then $\Phi_x(t)$ is a continuous distribution with the probability density, $\varphi_x(t)$, as given in (2). Analogously, the probability of y assuming a value between $-\infty$ and u independently of which value x assumes is given by

$$\Phi_y(u) = \Phi(u) = \int_{-\infty}^{u} du \int_{-\infty}^{\infty} \varphi(t, u)\, dt. \tag{3}$$

$\Phi_y(u)$ is called the **marginal distribution** of y, and under the same conditions as above it is a continuous distribution with the probability density

$$\varphi_y(u) = \int_{-\infty}^{\infty} \varphi(t, u)\, dt, \tag{4}$$

or briefly $\varphi(u)$. We wish to stress that although the marginal distributions are given uniquely by the two-dimensional distribution the converse is not true. For any given functions $\varphi_x(t) \geqq 0$ and $\varphi_y(u) \geqq 0$ satisfying only (4.4.3) we can always construct the two-dimensional distribution given by $\varphi(t, u) = \varphi_x(t)\varphi_y(u)$. However, as shown in the following example, there may exist infinitely many two-dimensional distributions, all having the same marginal distributions.

Example. For the normal distribution (4.12.7) the marginal distributions are, independently of ρ, also normal with the parameters μ_x, σ_x, and μ_y, σ_y, respectively. To show this it is convenient to put

$$x = \frac{t - \mu_x}{\sigma_x}, \qquad y = \frac{u - \mu_y}{\sigma_y}. \tag{5}$$

From (2) we then have, introducing (5) into (4.12.7),

$$d\Phi_x = \varphi_x(t)\,dt =$$

$$\frac{dx}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right]dy =$$

$$\frac{dx}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{x^2}{2}\right]\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2(1-\rho^2)}(y-\rho x)^2\right]dy =$$

$$\frac{dx}{\sqrt{2\pi}}\exp\left[-\frac{x^2}{2}\right]\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{v^2}{2}\right]dv,$$

if we introduce $v = \dfrac{y-\rho x}{\sqrt{1-\rho^2}}$ as a new variable. From the fact, proved in Chapter 7, that this last integral is equal to 1, and by introducing t instead of x (from (5)), we find

$$d\Phi_x = \varphi_x(t)\,dt = \frac{1}{\sqrt{2\pi}\,\sigma_x}\exp\left[-\frac{(t-\mu_x)^2}{2{\sigma_x}^2}\right]dt, \text{ Q.E.D.} \tag{6}$$

In the same way we find the corresponding result

$$d\Phi_y = \varphi_y(u)\,du = \frac{1}{\sqrt{2\pi}\,\sigma_y}\exp\left[-\frac{(u-\mu_y)^2}{2{\sigma_y}^2}\right]du. \tag{7}$$

Incidentally, (6) or (7) shows that the normalization factor in (4.12.7) is correct, i.e., that (4.12.5) is satisfied.

We define $\varphi_{x|y}(t|u)$, or briefly $\varphi(t|u)$, by

$$\varphi_{x|y}(t|u) = \frac{\varphi(t,\,u)}{\varphi_y(u)}, \tag{8}$$

and $\varphi_{y|x}(u|t)$, or briefly $\varphi(u|t)$, by

$$\varphi_{y|x}(u|t) = \frac{\varphi(t,\,u)}{\varphi_x(t)}. \tag{9}$$

Then we see from the identity

$$\varphi(t,\,u)\,dt\,du = \varphi_x(t)\,dt\,\varphi_{y|x}(u|t)\,du = \varphi_y(u)\,du\,\varphi_{x|y}(t|u)\,dt \tag{10}$$

and the multiplication law, V, that $d\Phi_{x|y} = \varphi_{x|y}(t|u)\,dt$ gives the conditioned probability of x assuming a value between t and $t+dt$ under the assumption that y has assumed a value between u and $u+du$. $d\Phi_{y|x} = \varphi_{y|x}(u|t)\,du$ has the corresponding interpretation. These two distributions are called the **conditional distributions.**

Exercise 1. Show that

$$\int_{-\infty}^{\infty} \varphi_{x|y}(t|u)\,dt = \int_{-\infty}^{\infty} \varphi_{y|x}(u|t)\,du = 1. \tag{11}$$

If $\varphi_{x|y}(t|u) = \varphi_x(t)$ for all values of t and u, which from (10) implies that $\varphi_{y|x}(u|t) = \varphi_y(u)$, we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ are **stochastically independent,** or simply **independent,** and we then have

$$\varphi(t, u) = \varphi_x(t)\varphi_y(u). \tag{12}$$

Exercise 2. Conversely, show that, if (12) holds true, $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

Exercise 3. By means of the results of the example show that for the normal distribution the two conditional distributions are also normal with the parameters

$\mu = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(u - \mu_y)$, $\sigma = \sqrt{1-\rho^2}\,\sigma_x$ and $\mu = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(t - \mu_x)$, $\sigma = \sqrt{1-\rho^2}\,\sigma_y$ respectively. That is,

$$d\Phi_{x|y} = \varphi_{x|y}(t|u)\,dt =$$

$$\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\,\sigma_x}\exp\left[-\frac{1}{2(1-\rho^2)\sigma_x^2}\left(t - \mu_x - \rho\frac{\sigma_x}{\sigma_y}(u-\mu_y)\right)^2\right]dt \tag{13}$$

and

$$d\Phi_{y|x} = \varphi_{y|x}(u|t)\,du =$$

$$\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\,\sigma_y}\exp\left[-\frac{1}{2(1-\rho^2)\sigma_y^2}\left(u - \mu_y - \rho\frac{\sigma_y}{\sigma_x}(t-\mu_x)\right)^2\right]du. \tag{14}$$

***§ 4.14.** By a function $\boldsymbol{w} = f(\boldsymbol{z}) = f(\boldsymbol{x}, \boldsymbol{y})$ we understand a new one-dimensional random variable $\boldsymbol{w}$ which assumes the value $f(t, u)$, when $\boldsymbol{x}$ assumes the value t and $\boldsymbol{y}$ the value u. The distribution function $\Phi_w(s)$ of $\boldsymbol{w}$, which is equal to the probability, $P(\boldsymbol{w} \leqq s)$, of $\boldsymbol{w}$ assuming a value smaller than or equal to s, is then from (4.12.1) given by the plane integral of $\varphi(t, u)$ over that region in the tu-plane in which $\boldsymbol{w} \leqq s$:

$$\Phi_w(s) = P(\boldsymbol{w} \leqq s) = \int\int_{w \leqq s} \varphi(t, u)\,dt\,du. \tag{1}$$

Often it is convenient, e.g., in evaluating (1), to express the distribution $d\Phi = \varphi(t, u)\,dt\,du$ in new variables t', u' given by $t = t(t', u')$, $u = u(t', u')$. It may be shown[1] that, if the so-called **Jacobian functional determinant**

$$\frac{\partial(t, u)}{\partial(t', u')} = \begin{vmatrix} \dfrac{\partial t}{\partial t'} & \dfrac{\partial t}{\partial u'} \\ \dfrac{\partial u}{\partial t'} & \dfrac{\partial u}{\partial u'} \end{vmatrix} \neq 0 \tag{2}$$

[1] See any textbook in analysis, e.g., that of R. Courant, *Differential and Integral Calculus,* New York, 1947.

in the whole tu-plane, we have

$$d\Phi = \varphi(t, u)\, dt\, du = \varphi(t(t', u'), u(t', u'))\cdot \left|\frac{\partial(t, u)}{\partial(t', u')}\right| dt'\, du' =$$

$$\varphi'(t', u')\, dt'\, du' = d\Phi', \qquad (3)$$

which is a direct generalization of (4.8.1).

Introducing into (1) the new variables $s = f(t, u)$ and $r = g(t, u)$ (the r being an arbitrary variable, e.g., $r = t$, for which (2) is satisfied) we find

$$\Phi_w(s) = \int_{-\infty}^{s} ds \int dr \varphi(t(r, s), u(r, s)) \left|\frac{\partial(t, u)}{\partial(r, s)}\right|. \qquad (4)$$

Differentiating (4) with respect to s we have that the probability density of w is given by

$$\varphi_w(s) = \int \varphi(t(r, s), u(r, s)) \left|\frac{\partial(t, u)}{\partial(r, s)}\right| dr. \qquad (5)$$

Example. Let x and y be independent and have the probability densities $\varphi_x(t)$ and $\varphi_y(u)$ respectively. For $w = x + y$ we then obtain from (5), putting $s = t + u$ and, e.g., $r = t$,

$$\varphi_w(s) = \int_{t+u=s} \varphi_x(t)\varphi_y(u) \left|\frac{\partial(t, u)}{\partial(r, s)}\right| dr = \int_{-\infty}^{\infty} \varphi_x(r)\varphi_y(s - r)\, dr \qquad (6)$$

since

$$\frac{\partial(t, u)}{\partial(r, s)} = 1.$$

We leave to the reader to generalize the formulae of § 4.9 to § 4.14 to random variables of more than two dimensions.

*§ 4.15. **Stochastic processes.** One often finds problems in which the distribution of a random variable depends on a non-random variable which is a continuously varying parameter such as time. In these cases we then speak of a **stochastic** or **random process.**[1] In Example 4, § 4.3, we have mentioned an example of a **discontinuous stochastic process,** i.e., a process in which the random variable is discontinuously distributed. In the theory of Brownian motions in which the position (x, y, z) of a Brownian particle is the three-dimensional random variable we have an example of a **continuous stochastic process,** i.e., one in which the random variable is continuously distributed. Here the total "probability mass," $P(V, t) = \int\int\int_V \varphi(x, y, z; t)\, dx\, dy\, dz$, inside a given volume V is a function of time. (In this and the two following topics we use the same

[1] For the general theory see, e.g., Khintchine, *Asymptotische Gesetze der Wahrscheinlichkeitsrechnung;* Feller, *Math. Ann.*, **113**, 113, 1937.

letter for the random variable and the corresponding variable in the distribution function.) Since probability is conserved, i.e., can neither disappear nor be created, any change in P must correspond to a flow of "probability mass" through the boundary surface F of V, a "probability current" in the mechanical picture of §4.6. If the vector $\vec{s}$ denotes the **probability current density**, i.e., the probability mass crossing a unit area perpendicular to the direction of $\vec{s}$ per unit time, this conservation law becomes

$$-\frac{dP}{dt} = -\frac{d}{dt}\int\int\int_V \varphi\, dv = \int\int_F s_n\, df, \tag{1}$$

in which dv is a volume element, df a surface element with normal $\vec{n}$, positive outwards, and s_n the component of $\vec{s}$ along $\vec{n}$. By means of Gauss's theorem we obtain from (1) the so-called **continuity equation**

$$\operatorname{div} \vec{s} + \frac{\partial \varphi}{\partial t} = \frac{\partial s_x}{\partial x} + \frac{\partial s_y}{\partial y} + \frac{\partial s_z}{\partial z} + \frac{\partial \varphi}{\partial t} = 0, \tag{2}$$

which is analogous to that found in electricity, heat conduction, and quantum theory.

Often it may be a good approximation to assume that $\vec{s}$ is proportional to the gradient of φ, as is the case, for example, in heat conduction. Then

$$\vec{s} = -D \operatorname{grad} \varphi = -D\left(\frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y}, \frac{\partial \varphi}{\partial z}\right), \tag{3}$$

in which $D(>0)$ is a constant called the *diffusion coefficient*. Now (2) reduces to

$$D \operatorname{div} \operatorname{grad} \varphi = D\,\Delta\varphi = \frac{\partial \varphi}{\partial t}, \tag{4}$$

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Exercise. Show that the normal distribution given by

$$d\Phi = \varphi(x, y, z; t)\, dv = \frac{1}{2\sqrt{\pi D t}} \exp\left[-\frac{(x-\mu_x)^2 + (y-\mu_y)^2 + (z-\mu_z)^2}{4Dt}\right] dx\, dy\, dz, \tag{5}$$

which is the three-dimensional generalization of (4.12.6) for $\sigma_x = \sigma_y = \sigma_z = \sqrt{2Dt}$, is a solution of (4). For $t \to 0$, it is seen that $\Phi \to \epsilon(x-\mu_x)\epsilon(y-\mu_y)\epsilon(z-\mu_z)$ (cf. Example 1, §4.3), i.e., that (5) corresponds to the initial condition that the Brownian particle is at the point (μ_x, μ_y, μ_z) with certainty for $t = 0$. Using the Dirac δ-function (cf. the end of §4.4) we may also express this by writing

$$\varphi(x, y, z; t = 0)\, dv = \delta(x-\mu_x)\delta(y-\mu_y)\delta(z-\mu_z)\, dv.$$

If, besides the "diffusion current" in (3), we also have a "convection current" with the velocity $\vec{v}$, we have to add in (3) a term $\varphi \vec{v}$. Furthermore, if we allow D to be a function of (x, y, z), (2) instead of (4) then assumes the general form

$$\Delta(D\varphi) - \operatorname{div}(\varphi \vec{v}) = \frac{\partial \varphi}{\partial t}. \tag{6}$$

This so-called **Planck-Fokker equation** is the general equation of a three-dimensional continuous stochastic process; it determines φ uniquely for all times if φ is known for one time, $t = t_0$. (It may be generalized to any number of dimensions.)

Stochastic processes, both discontinuous and continuous, play an ever-increasing role in many practical applications such as physics, engineering (especially telephony), biology, and insurance.[1]

*§4.16. **Statistical mechanics.** Special cases of continuous stochastic processes and many-dimensional distributions are found in **statistical mechanics**, that branch of mechanics which studies the motion of bodies when the initial conditions are not sufficiently well known to permit the causal description that is in principle always possible in classical mechanics.[2] The exact characterization of the state of a mechanical system with f degrees of freedom consists of stating the "point" $\vec{r} = (q_1, \cdots, q_f, p_1, \cdots, p_f)$ in the $2f$-dimensional *phase space*, i.e., the numerical values of the f generalized coordinates, $q_1, \cdots, q_f$ and the corresponding f generalized momenta, $p_1, \cdots, p_f$ (which are generalizations of cartesian coordinates and cartesian momenta, respectively). Since in more complicated cases, such as in describing 1 gram of hydrogen (cf. §1.2), it is either practically impossible or inconvenient to measure, and state, all these $2f$ quantities, we use a statistical description; i.e., we treat our vector as a $2f$-dimensional random variable with a continuous distribution function which depends on time. Thus

$$d\Phi = \varphi(\vec{r}, t)\, dv = \varphi(q_1, \cdots, q_f, p_1, \cdots, p_f; t)\, dq_1 \cdots dq_f \cdot dp_1 \cdots dp_f, \tag{1}$$

where dv is the element of volume in phase space. It may be shown that φ satisfies a special form of the Planck-Fokker equation (4.15.6) with $\vec{v} = \left(\frac{dq_1}{dt}, \cdots, \frac{dp_f}{dt}\right)$ being the "velocity" of the phase point $\vec{r}$:

[1] Among the numerous applications we shall mention only: Arley, *Stochastic Processes and Cosmic Radiation* (theory of discontinuous processes and application to the theory of cosmic radiation); Chandrasekhar, *Stochastic Problems in Physics and Astronomy;* Lundberg, *Random Processes and Sickness and Accident Statistics* (theory of discontinuous processes and application to insurance); Bartlett, *Stochastic Processes* (general theory with illustrative examples from various fields).

[2] Cf., e.g., Gibbs, *Elementary Principles in Statistical Mechanics*, Yale, 1903 (general theory); Tolman, *The Principles of Statistical Mechanics*, Oxford, 1938 (general theory); Lindsay, *Introduction to Physical Statistics*, Wiley, 1941 (many applications). The subject is treated from the point of view of modern probability in Khinchin, *Mathematical Foundations of Statistical Mechanics*, Dover Publications, 1949, and in Moyal, Stochastic Processes and Statistical Physics, *J. Roy. Stat. Soc.*, B, vol. 11, 1949. See also Born, *Cauce and Chance*, Oxford, 1949.

$$\operatorname{div}(\varphi \vec{v}) + \frac{\partial \varphi}{\partial t} = \sum_{k=1}^{f} \left(\frac{\partial}{\partial q_k} \left(\varphi \frac{dq_k}{dt} \right) + \frac{\partial}{\partial p_k} \left(\varphi \frac{dp_k}{dt} \right) \right) + \frac{\partial \varphi}{\partial t} = 0, \tag{2}$$

which determines φ uniquely for all times if φ is known for one time, $t = t_0$.

We note that the difference between the general stochastic process described in (4.15.6) and the special one described in (2) is that the randomness enters in two different ways. In the former case it is the transition from the initial state, $\vec{r}_1 = (x_1, y_1, z_1)$, of the system at a time t_1 to the final state, $\vec{r}_2 = (x_2, y_2, z_2)$, at a time $t_2 (> t_1)$, which is random, whereas in the latter case it is the initial state itself which is random, the transition to the final state being causally (uniquely) determined from Newton's mechanical laws if the initial state is known to be exactly $\vec{r}' = (q_1', \cdots, p_f')$, i.e., if the initial distribution is a causal distribution: $\Phi = \epsilon(q_1 - q_1') \cdots \epsilon(p_f - p_f')$.

If the temperature of our system is kept constant, experience shows that φ will approach for $t \to \infty$ a certain *stationary distribution* which is independent of the initial state; we say that our system has come into *thermal* or *statistical equilibrium*. Such equilibria are the main subject studied in the applications of statistical mechanics. However, recently the problem of how the equilibria are reached, the so-called *transport phenomena*, has also been studied. Mathematically this means that our system is considered as a small part of a very large system (heat bath). It can then be shown that the stationary distribution, Φ_{stat}, is given by Gibbs' **canonical distribution,** $E = E(q_1, \cdots, p_f)$ being the total energy,

$$d\Phi_{\text{stat}} = \varphi_{\text{stat}}(\vec{r})\, dv = c \exp\left[-\frac{E(q_1, \cdots, p_f)}{kT} \right] \cdot dq_1 \cdots dq_f \cdot dp_1 \cdots dp_f. \tag{3}$$

Here c is a normalization constant often written $c = \exp\left[\frac{\psi}{kT}\right]$, T is the absolute temperature, and k is a physical constant called *Boltzmann's constant*. Thus $\theta = kT (> 0)$ is an arbitrary parameter called the *modulus* of the distribution. (In particular, if we consider only such variations of $q_1, \cdots, p_f$ in phase space for which E = constant, we call the distribution obtained from (3) the **microcanonical distribution.**)

Exercise. Let the system consist of N particles and be conservative; i.e., in cartesian coordinates $E = T + V = \sum_{i=1}^{N} \frac{1}{2m_i} (p_{xi}^2 + p_{yi}^2 + p_{zi}^2) + V(x_1, \cdots z_N)$, T being the kinetic, V the potential energy, and $p_{xi} = m_i v_{xi}$, and so forth. Show that the components of the velocities are independent of the coordinates, are mutually independent, and that each has a marginal distribution which is normal with parameters $\mu = 0$, $\sigma = \sqrt{kT/m_i}$: Maxwell-Boltzmann's law. (For the distribution of $v_i = \sqrt{v_{xi}^2 + v_{yi}^2 + v_{zi}^2}$ cf. Exercise 3, §7.4.)

Example 1. We want to remind the reader how (3) should be compared with experience (cf. Chapter 1). To that purpose we had to consider a large number, n, of identical copies of the mechanical system considered, all being in thermal

equilibrium at the same absolute temperature T. Let us assume that, at one and the same time, we had measured the state of each of these n systems, i.e., their phase points, and found the results $\vec{r}_1 = (q_{11}, \cdots, p_{f1}), \cdots, \vec{r}_n = (q_{1n}, \cdots, p_{fn})$. Let V be an arbitrary region in phase space, and let n_V be the number among our n measured phase points which lie inside V. Then the relative frequency n_V/n would be the experimental value of the theoretical probability $P = \int_V \cdots \int \varphi \, dv$ of finding $\vec{r}$ inside V.[1] This empirical population of n measured phase points is obviously of the type (3) in the classification of § 1.1; we shall call it the empirical *space* population. Now we could also construct another empirical population of n phase points by considering only one copy of our mechanical system, assumed to be in thermal equilibrium at the absolute temperature T, but then measuring its state, i.e., the phase point $\vec{r}$, at a large number of consecutive times $t_1 < t_2 < \cdots < t_n$ throughout a long time interval. This second empirical population of n measured phase points is obviously of the type (2) of § 1.1; we shall call it the empirical *time* population. If, now, we calculate the relative frequency n_V'/n in this second empirical population, experience shows, indirectly, that n_V'/n may also be regarded as an experimental value of the same probability $P = \int_V \cdots \int \varphi \, dv$. This remarkable experimental finding that both a space and a time population lead to the same results is a characteristic feature of systems in *equilibrium*. However, on reflection, this is not surprising; if the system is in equilibrium it is a natural conjecture that during a sufficiently long time its phase point $\vec{r} = \vec{r}(t)$ will move in phase space in such a way that it passes arbitrarily close to each phase point and will spend a time in any region of phase space, V, approximately proportional to the probability $P = \int_V \cdots \int \varphi \, dv$. In fact, in advanced probability, this so called **ergodic theorem** may be proved under very general conditions.[2]

Example 2. From the canonical distribution (3) of the phase point of our system the distribution of any other physical quantity in the system can be deduced. Thus, if our system consists of N identical particles, such as a gas containing N identical molecules, i.e., $f = 3N$, we may for instance ask for the following. Instead of considering $\vec{r} = (q_1, \cdots, q_f, p_1, \cdots, p_f)$ as one $2f(=6N)$-dimensional random variable, i.e., considering one phase point in a $6N$-dimensional phase space, called the *γ-space*, we may also consider $\vec{r}$ as being N, in general dependent, random variables, each of 6 dimensions; i.e., we consider N phase points in one 6-dimensional phase space, called *μ-space*. Now if we divide this

[1] It should be mentioned that in reality n_V/n cannot, of course, be measured directly, since not even a single phase point can actually be measured, because of the enormous value of $f \sim 10^{23}$. However, other properties of the system can be measured, and thus (3) can be tested indirectly.

[2] See, e.g., E. Hopf, Ergodentheorie, *Ergeb. d. Math.*, vol. 5, Heft 2, Berlin, 1937.

μ-space into a finite number, k, of regions, $\omega_1, \cdots, \omega_k$, called *cells*, we may ask for the probability of a certain *cell distribution*, i.e., that among the N phase points there are just N_1 in ω_1, N_2 in ω_2, $\cdots$, N_k in ω_k. Since $N_1, \cdots, N_k$ may assume different values in different measurements (performed either in a space or a time population), $N_1, \cdots, N_k$ will form a k-dimensional random variable. This cell distribution may itself be described by a distribution function $F(p)$, equal to the relative number of phase points lying in cells with a number smaller than or equal to $p = 1, 2, \cdots, k$. Formally, the function $F(p)$ is a step function having exactly the same properties as a discontinuous distribution function (§4.3). Thus we see that in certain cases a whole distribution function may itself be treated as a random variable (with a number of dimensions equal to the number of discontinuities, viz., k in our case). If we choose k very large, $F(p)$ may be approximated by a continuous distribution function. Thus we are led to consider a whole continuous distribution function $F(t)$, i.e., each value of $F(t)$, as a random variable and thus generally to speak of the distribution of a distribution. However, the dimensions of this random variable, being an element in "function space," is obviously highly infinite. Hence we are outside the realm of classical probability, but in the modern, very general, theory of Kolmogoroff such infinite-dimensional probabilities may also be treated.

The cell distribution in μ-space must not be confused with the canonical distribution in γ-space; the former can always, both in classical and in quantum statistics, be deduced from the latter, but for the further discussion of these questions and of the further application of probability outlined in this topic we must, however, refer to textbooks in statistical mechanics.

***§4.17. Quantum theory.** In quantum theory, which may be said to be one large theory of a special type of stochastic processes, these stochastic processes are described differently from those mentioned in §4.15 and §4.16.[1] Whereas in diffusion phenomena the randomness enters only in the transition from one state of the system to another and in statistical mechanics it enters only in the initial state, the transition itself being causal, in quantum theory the randomness enters both in the initial state and in the transition from this state to another. A further difference is that in statistical mechanics the initial state is considered random only for practical reasons, an exact measurement being in principle possible, while in quantum theory it is random by principle since the interaction between the measured system and the measuring apparatus cannot be neglected for atomic phenomena (cf. §1.2). Thus the initial state cannot be measured accurately enough to permit a causal description. Any state of a system can, therefore, be described only by a certain distribution function, which in general is different from the causal one, $\epsilon(t - \mu)$. The transition from classical to quantum theory, both in mechanics and in electrodynamics, takes place by replacing, in a certain way called *quantization*, the causal distributions of the exactly measurable physical quantities of classical theory by more general distributions. These general distributions contain the classical, causal ones as special limiting cases when $h \to 0$, h being a new fundamental constant, called *Planck's constant*. Instead of giving these distributions directly by their distribution functions, the quantum theory first gives certain complex functions called **probability amplitudes**, the squares of the numerical values of which are equal to the differential distribution functions.

[1] For a discussion of fundamental experiments and principles, as well as mathematical formalism, see, e.g., W. Heisenberg, *The Physical Principles of the Quantum Theory*, Chicago, 1930.

Thus, for a single particle of mass m moving in a conservative field of force with potential energy $V(x, y, z)$, any state is characterized by a so-called **Schroedinger wave function,** or probability amplitude for position, $\psi(x, y, z; t)$, satisfying Schroedinger's wave equation which is analogous to the diffusion equation (4.15.4)

$$\Delta\psi - \frac{8\pi^2 m}{h^2}\psi = -i\frac{4\pi m}{h}\frac{\partial\psi}{\partial t}, \tag{1}$$

in which Δ is defined as in (4.15.4), h is Planck's constant, and $i = \sqrt{-1}$. Here $|\psi|^2$ is the probability density at the time t of the distribution of the position (x, y, z) of the particle; i.e., the probability of finding it inside any region V of ordinary space at the time t is

$$P(V; t) = \int\int\int_V |\psi(x, y, z; t)|^2\, dv; \qquad \int\int\int |\psi|^2\, dv = 1. \tag{2}$$

Exercise. Deduce from (1) the continuity equation corresponding to (4.15.2)

$$\operatorname{div} \vec{s} + \frac{\partial}{\partial t}|\psi|^2 = 0, \quad \vec{s} = \frac{h}{4\pi m i}(\psi^* \operatorname{grad} \psi - \psi \operatorname{grad} \psi^*), \tag{3}$$

in which * means complex conjugate and grad ψ is defined as in (4.15.3). (Multiply (1) by ψ^*, the corresponding equation for ψ^* by ψ, subtract, and use partial integration.)

The fact that it is the probability amplitudes and not the distribution functions themselves which enter primarily in the mathematical description gives rise to interference effects whereby the usual probability laws are somewhat modified (see, e.g., Heisenberg *The Physical Principles of the Quantum Theory*, Chapter IV, §2). From the ψ-function characterizing a certain state of the mechanical system considered the distribution of any physical quantity can be deduced. For a further discussion of this application of probability we must refer to textbooks in quantum theory.

5.

MEAN VALUE AND DISPERSION

§ **5.1.** For an accurate characterization of a distribution it is, of course, necessary to know the whole distribution function $\Phi(t)$ or, for a continuous distribution, the whole probability density $\varphi(t)$ of x. Now, in most practical applications the probability mass of x will be concentrated mainly within a relatively narrow interval, and therefore to get a rough idea of the whole distribution it is often appropriate to indicate the position of this interval by some **measure of location,** giving a typical value of x. Although any such measure is, as a rule, uniquely determined by $\Phi(t)$, the converse is obviously not true.

There may be constructed infinitely many such measures of location, but in practice the following three are those mostly used: the **mode** is defined as the most probable value of x, i.e., for discontinuous and continuous distributions that value of t for which φ_i and $\varphi(t)$ respectively are maxima; the **median** is defined as that value of t for which $\Phi(t) = \frac{1}{2}$, i.e., for which the probability of x assuming a value smaller than t is equal to the probability of x assuming a value larger than t; the **mean value,** or briefly the **mean,** is in the mechanical picture of § 4.6 defined as the center of gravity of the whole probability mass. Which of these, or of other, measures of location to use is quite arbitrary, it being solely a question of convenience. However, since most rules of calculation are simpler for means, this measure of location is that most commonly used, but in special problems other measures may be more convenient.

Exercise 1. Discuss whether or not the mode, the median, and the mean always exist and are uniquely determined by $\Phi(t)$ (treat the discontinuous and the continuous cases separately). Next show that, if $\varphi(t)$ exists and is symmetric about $t = a$, then the median and the mean are equal to a; and that, furthermore, if $\varphi(t)$ has only one maximum, the mode is also equal to a.

From the mechanical picture of § 4.6 and the definition of the center of gravity of a body it follows that the mean of x, which we shall denote by either $\mathfrak{M}\{x\}$ or μ_x or briefly μ,[1] is, if it exists, defined by

[1] Other symbols and names used are $\langle x \rangle$, or $\langle x \rangle_{\text{av}}$ or av x from "**average** of x"; $E\{x\}$, from "**expectation value** of x," a term originating from the applica-

$$\mu = \mathfrak{M}\{x\} = \frac{\sum_{i=-\infty}^{\infty} t_i\varphi_i}{\sum_{i=-\infty}^{\infty} \varphi_i}, \qquad \mu = \mathfrak{M}\{x\} = \frac{\int_{-\infty}^{\infty} t\varphi(t)\,dt}{\int_{-\infty}^{\infty} \varphi(t)\,dt} \tag{1}$$

for discontinuous and continuous distributions, respectively (t_i and φ_i are defined in §4.3, $\varphi(t)$ in §4.4). Thus for calculating the mean we need not normalize to one. However, since we have assumed that this is always the case, i.e., (4.3.2) or (4.4.3) to be satisfied, we have

$$\mu = \mathfrak{M}\{x\} = \begin{cases} \sum_{i=-\infty}^{\infty} t_i\varphi_i & \text{for discontinuous distributions} \\ \int_{-\infty}^{\infty} t\varphi(t)\,dt & \text{for continuous distributions.} \end{cases} \tag{2}$$

(In general, it is, furthermore, assumed that the sum or integral is absolutely convergent if μ exists at all.)

Example 1. If x is *bounded*, i.e., x can assume only values lying in a finite interval $g \leqq t \leqq G$, we have in the discontinuous case

$$g = g\sum_i \varphi_i \leqq \sum_i t_i\varphi_i = \mu \leqq G\sum_i \varphi_i = G$$

and in the continuous case

$$g = g\int_g^G \varphi(t)\,dt < \int_g^G t\varphi(t)\,dt = \mu < G\int_g^G \varphi(t)\,dt = G.$$

In both cases μ lies between the smallest and largest values that x can assume, which fact justifies the name mean.

Example 2. If x can assume only the values 1 and 0 with the probability θ and $1 - \theta$ respectively, we have

$$\mu = \mathfrak{M}\{x\} = 1 \cdot \theta + 0 \cdot (1 - \theta) = \theta; \tag{3}$$

i.e., the mean is equal to the probability of x assuming the value 1.

---

tions of probability to games of chance; $\bar{x}$ which is perhaps the most convenient symbol and is much used, e.g., in physics. However, in statistics $\bar{x}$ denotes something different (cf. p. 76), and therefore we shall not use it to denote means. The reason for not writing $\mathfrak{M}(x)$ is that $\mathfrak{M}\{x\}$ is not a function of x but a "functional," i.e., a number associated to the whole distribution function $\Phi(t)$ of x. French, **valeur moyenne** (or **espérance mathématique**); German, **Mittelwert** (or **Erwartungswert**).

Example 3. If x assumes with equal probability the values 1, 2, $\cdots$, ν, we have

$$\mu = \mathfrak{M}\{x\} = 1 \cdot \frac{1}{\nu} + 2 \cdot \frac{1}{\nu} + \cdots + \nu \cdot \frac{1}{\nu} = \frac{\nu(\nu+1)}{2}\frac{1}{\nu} = \frac{\nu+1}{2}. \quad (4)$$

Thus, if x is the result of throwing a die, we have $\nu = 6$, i.e., $\mu = 3.5$.

Example 4. For the binomial distribution (Example 2, § 4.3) we have by means of the binomial theorem

$$\mu = \mathfrak{M}\{x\} = \sum_{i=0}^{\nu} i \binom{\nu}{i} \theta^i (1-\theta)^{\nu-i} =$$

$$\nu\theta \sum_{i=1}^{\nu} \binom{\nu-1}{i-1} \theta^{i-1}(1-\theta)^{(\nu-1)-(i-1)} =$$

$$\nu\theta \sum_{i'=0}^{\nu-1} \binom{\nu-1}{i'} \theta^{i'}(1-\theta)^{(\nu-1)-i'} = \nu\theta. \quad (5)$$

We shall later give a much simpler proof of this formula (cf. Example 1, § 6.4).

Example 5. For Poisson's distribution (Example 3, § 4.3) we have

$$\mathfrak{M}\{x\} = \sum_{i=0}^{\infty} i e^{-\mu} \frac{\mu^i}{i!} = e^{-\mu}\mu \sum_{i=1}^{\infty} \frac{\mu^{i-1}}{(i-1)!} = \mu, \quad (6)$$

which shows the reason for using the letter μ for the parameter of this distribution. Thus in Example 4, § 4.3, we have $\mu = \lambda t$; i.e., λ is the mean number of calls per unit of time.

***Example 6.** The result of the preceding example has an important application to the theory of Geiger-Müller counters (cf. Example 5, § 4.3). Let n Geiger-Müller counters all have the same resolving time τ, which is the minimum time interval between two consecutive passages which the counter can distinguish as two separate events. Next let the mean number of counts per unit time for the n counters be $N_1, \cdots, N_n$, respectively. We ask for the mean number of spurious, or accidental, coincidences per unit time between the n counters. First, the conditioned probability of each of the other $n-1$ counters being struck at least once within the time interval τ after one of them, say number 1, has been struck is

$$(1 - e^{-N_2\tau}) \cdots (1 - e^{-N_n\tau})$$

(why?). Therefore the mean number of accidental coincidences in which counter 1 is struck first is N_1 times this probability. Finally, the counter struck first may either be number 1 or number 2, $\cdots$, or number n. Thus the total mean number of accidental coincidences per unit time is, assuming $N_1\tau \ll 1, \cdots, N_n\tau \ll 1$,

$$nN_1 \cdots N_n\tau^{n-1}. \tag{7}$$

Example 7. Also for Pascal's distribution (Example 6, §4.3) the parameter μ is so chosen that it is equal to the mean value:

$$\mathfrak{M}\{x\} = \sum_{i=0}^{\infty} i \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^i = \mu. \tag{8}$$

This may be shown by the following device, which is often useful. If we put $\frac{\mu}{1+\mu} = x(<1)$, we have

$$\sum_{i=0}^{\infty} ix^i = x\sum_{i=0}^{\infty} ix^{i-1} = x\frac{d}{dx}\sum_{i=0}^{\infty} x^i = x\frac{d}{dx}\frac{1}{1-x} = \frac{x}{(1-x)^2}.$$

Inserting this in (8) and introducing μ instead of x the result follows at once.

Exercise 2. For the causal distribution (Example 1, §4.3), for Pólya's distribution (Example 7, §4.3), and for Laplace's distribution (Example 5, §4.4), show that the parameter μ is also equal to the mean value.

Example 8. We shall see later (§7.3) that for the normal distribution (Example 3, §4.4) the parameter μ is again just the mean value, which fact also follows from the facts mentioned in Exercise 1.

Example 9. For Cauchy's distribution (Example 4, §4.4) we have

$$\mathfrak{M}\{x\} = \frac{1}{\pi\alpha}\int_{-\infty}^{\infty} t\frac{dt}{1+\frac{(t-\mu)^2}{\alpha^2}} = \int_{-\infty}^{\infty}((t-\mu)+\mu)\varphi(t)\,dt =$$

$$\frac{\alpha}{2\pi}\left|\ln\left(1+\frac{(t-\mu)^2}{\alpha^2}\right)\right|_{-\infty}^{\infty} + \mu = \infty - \infty + \mu.$$

In this case the expression for the mean value is undetermined. From the symmetry of $\varphi(t)$ about $t = \mu$ it is, however, natural to define the conditionally convergent integral so that

$$\mathfrak{M}\{x\} = \mu \tag{9}$$

(which is called the Cauchy principal value). It will be seen that both the mode and the median exist for this distribution and are equal to the parameter μ (cf. Exercise 1).

Exercise 3. Show that from the distribution of Example 2, § 4.4, the mean life time of a radioactive atom is (cf. Appendix 1)

$$\mathfrak{M}\{x\} = \int_0^\infty te^{-\lambda t}\lambda\, dt = \frac{1}{\lambda}\left(= \frac{T}{\ln 2}\right), \tag{10}$$

where T is the half lifetime (cf. the example, § 3.7). In example 4, § 4.3, $1/\lambda$ thus gives the mean time between two consecutive calls (cf. Example 2, § 4.4), which is, therefore, the reciprocal of the mean number of calls per unit time as found in Example 5.

§ 5.2. Quite generally we define $\mathfrak{M}\{y\}$, where $y = f(x)$, cf. § 4.8, by

$$\mathfrak{M}\{y\} = \mathfrak{M}\{f(x)\} = \begin{cases} \sum_{i=-\infty}^{\infty} f(t_i)\varphi_i & \text{for } x \text{ discontinuous} \\ \int_{-\infty}^{\infty} f(t)\varphi(t)\, dt & \text{for } x \text{ continuous} \end{cases} \tag{1}$$

(if these expressions exist and are absolutely convergent). For it can be proved that (1) always gives the same number that would be obtained from (5.1.2) if we had first worked out the distribution of y from that of x.[1]

Exercise 1. Prove this under the special conditions of § 4.8, using (4.8.1).

Exercise 2. Let x be the result of throwing a die, and let

$$y = f(x) = \begin{cases} 0 & \text{for } x \text{ even} \\ 1 & \text{for } x \text{ odd.} \end{cases}$$

Find the distribution of y and show that both (1) and (5.1.2) give the mean value, ½.

Exercise 3. Show that for $y = a$, a being a constant, we have

$$\mathfrak{M}\{a\} = a. \tag{2}$$

Next show that for $y = ax$ we have

$$\mathfrak{M}\{ax\} = a\mathfrak{M}\{x\}. \tag{3}$$

If we put $y = y_1 + y_2$, where y_1 and y_2 are two arbitrary functions of the *same* random variable x, (1) gives

$$\mathfrak{M}\{y_1 + y_2\} = \mathfrak{M}\{y_1\} + \mathfrak{M}\{y_2\}. \tag{4}$$

In § 6.1 we shall prove that this holds true generally.

[1] See, e.g., Cramér, *Random Variables*, p. 19.

From (4) and (2) we get as a special result

$$\mathfrak{M}\{\boldsymbol{x} + a\} = \mathfrak{M}\{\boldsymbol{x}\} + a. \tag{5}$$

We stress that unless $f(\boldsymbol{x})$ is a linear function the mean is *not* invariant by the transformation from $\boldsymbol{x}$ to $\boldsymbol{y}$ given by $\boldsymbol{y} = f(\boldsymbol{x})$; i.e.,

$$\mathfrak{M}\{f(\boldsymbol{x})\} \neq f(\mathfrak{M}\{\boldsymbol{x}\}). \tag{6}$$

For example, we see that, in Exercise 2, $f(\mathfrak{M}\{\boldsymbol{x}\})$ is not even defined. (Cf. § 6.5.)

***Example 1.** In the Maxwell-Boltzmann velocity distribution (Example, § 4.8) we find by means of Appendix 1

$$\mathfrak{M}\{\boldsymbol{v}\} = \frac{\int_0^\infty \alpha t^3 \exp[-\beta t^2]\,dt}{\int_0^\infty \alpha t^2 \exp[-\beta t^2]\,dt} = \frac{\beta^{3/2}}{\beta^{1/2}} \frac{\int_0^\infty x^3 \exp[-x^2]\,dx}{\int_0^\infty x^2 \exp[-x^2]\,dx} = \frac{2}{\sqrt{\pi\beta}} = 2\sqrt{\frac{2kT}{\pi m}}, \tag{7}$$

and for the corresponding energy distribution

$$\mathfrak{M}\{\boldsymbol{E}\} = \frac{\int_0^\infty \alpha' u^{3/2} \exp[-\beta' u]\,du}{\int_0^\infty \alpha' u^{1/2} \exp[-\beta' u]\,du} = \frac{\beta'^{3/2}}{\beta'^{1/2}} \frac{\int_0^\infty x^{3/2} e^{-x}\,dx}{\int_0^\infty x^{1/2} e^{-x}\,dx} = \frac{3}{2\beta'} = \frac{3}{2}kT. \tag{8}$$

Thus $\mathfrak{M}\{\boldsymbol{E}\} = \frac{3}{2}kT$ is *not* equal to $\frac{1}{2}m(\mathfrak{M}\{\boldsymbol{v}\})^2 = (4/\pi)kT$. We note that this example shows that it is not always necessary to calculate the normalization factor α, namely, if we are interested only in mean values.

The last example demonstrates one serious drawback of the mean value as a measure of location, viz., that it is not invariant for non-linear transformations. The median is, therefore, sometimes preferred for the mean, since under the conditions of § 4.8 the median of a function is that function of the median.

Exercise 4. Verify this.

***Example 2.** A number of special expressions for the function in (1) are often used in the literature (for simplicity we consider here only continuous distributions, and we use x as the letter for the integration

variable). For $y = x^k$, k a real number $\geqq 0$, (1) gives

$$\mathfrak{M}\{x^k\} = \int_{-\infty}^{\infty} x^k \varphi(x)\, dx = \mu_k, \tag{9}$$

which, if it exists, is called x's **moment of order** k (with respect to the point $x = 0$). Analogously, $\mathfrak{M}\{|x|^k\}$ is called x's **absolute moment of order** k, and $\mathfrak{M}\{(x - \mu)^k\}$ is called x's **central moment of order** k. It is seen that μ_k is a generalization of $\mu_1 = \mu$; furthermore, that, if it exists, it is uniquely determined by $\varphi(x)$. Conversely, under certain special conditions, $\varphi(x)$ is uniquely determined by its integer moments $\mu_1, \mu_2, \cdots$.[1]

For $y = x^{(k)} = x(x - 1) \cdots (x - k + 1)$, $k = 1, 2, 3 \cdots$, (1) gives

$$\mathfrak{M}\{x^{(k)}\} = \int_{-\infty}^{\infty} x(x - 1) \cdots (x - k + 1)\varphi(x)\, dx = \mu_{(k)}, \tag{10}$$

which, if it exists, is called x's **factorial moment of order** k.

For $y = t^x$, t a real number $\geqq 0$, (1) gives

$$\mathfrak{M}\{t^x\} = \int_{-\infty}^{\infty} t^x \varphi(x)\, dx = \gamma_x(t), \tag{11}$$

which, if it exists, is a function of t called x's **generating function.**

For $y = e^{tx}$, t a real number, (1) gives

$$\mathfrak{M}\{e^{tx}\} = \int_{-\infty}^{\infty} e^{tx} \varphi(x)\, dx = \mu_x(t), \tag{12}$$

which, if it exists, is a function of t called x's **moment generating function,** since, if $\mu_x(t)$ may be expanded in a power series in t, the kth coefficient is obviously $\mu_k/k!$.

For $y = e^{itx}$, t a real number, $i = \sqrt{-1}$, (1) gives

$$\mathfrak{M}\{e^{itx}\} = \int_{-\infty}^{\infty} e^{itx} \varphi(x)\, dx = \chi_x(t), \tag{13}$$

which function of t is seen to exist for all values of t and for all distributions. It is called x's **characteristic function,** and it is a very useful tool in modern probability.[2] It is seen that, if it may be expanded in a power series of it, the kth coefficient is $\mu_k/k!$. Not only is $\chi_x(t)$ obviously uniquely determined by $\Phi(t)$, but, conversely, it determines $\Phi(t)$ uniquely: for a continuous distribution we have

[1] See, e.g., Cramér, *Mathematical Methods of Statistics*, p. 176.

[2] For proofs of the following statements and applications, see, e.g., Cramér, *Mathematical Methods of Statistics*, or Cramér, *Random Variables*, Chapter IV.

from the theory of Fourier integrals

$$\varphi(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \chi(t)\, dt, \tag{14}$$

and for a discontinuous distribution (cf. Exercise 5) we have, from the theory of so-called almost periodic functions,

$$\varphi_j = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-itx_j} \chi(t)\, dt, \tag{15}$$

in which x_j is the jth step-point, φ_j the jth step of $\Phi(t)$ (cf. §4.3; if x_j is not one of the step-points, (15) gives directly 0). Furthermore, we have the important theorem that, if $\Phi_1(t)$, $\Phi_2(t)$, $\cdots$ is a series of distribution functions and $\chi_1(t)$, $\chi_2(t)$, $\cdots$ the corresponding series of characteristic functions, then the necessary and sufficient condition for the convergence of $\Phi_n(t)$ towards a distribution function $\Phi(t)$ is that, for every t, $\chi_n(t)$ converges to a limit, $\chi(t)$, which is continuous at $t = 0$. $\chi(t)$ is then the characteristic function of $\Phi(t)$.

***Exercise 5.** Write down the expressions corresponding to (9)–(13) for discontinuous distributions.

***Exercise 6.** Show that, if x has the characteristic function $\chi_x(t)$, then $y = ax + b$, where a (> 0) and b are arbitrary constants, has the characteristic function $\chi_y(t) = \exp[bit]\chi_x(at)$.

***Exercise 7.** Show that the characteristic function is

$$\chi(t) = e^{i\mu t} \qquad \text{for the causal distribution} \tag{16}$$

$$\chi(t) = (1 + \theta(e^{it} - 1))^{\nu} \qquad \text{for the binomial distribution} \tag{17}$$

$$\chi(t) = \exp[\mu(e^{it} - 1)] \qquad \text{for Poisson's distribution} \tag{18}$$

$$\chi(t) = (1 + \mu(1 - e^{it}))^{-1} \qquad \text{for Pascal's distribution} \tag{19}$$

$$\chi(t) = (1 + \beta\mu(1 - e^{it}))^{-1/\beta} \qquad \text{for Pólya's distribution} \tag{20}$$

$$\chi(t) = \exp\left[i\mu t - \frac{\sigma^2}{2} t^2\right] \qquad \text{for the normal distribution} \tag{21}$$

$$\chi(t) = \exp[i\mu t - \alpha|t|] \qquad \text{for Cauchy's distribution} \tag{22}$$

$$\chi(t) = \frac{\exp[i\mu t]}{1 + \alpha^2 t^2} \qquad \text{for Laplace's distribution.} \tag{23}$$

Next find the corresponding generating and moment generating functions.

***Exercise 8.** Expanding the natural logarithm of $\chi(t)$ into a power series of it and calling the coefficients $\kappa_k/k!$, κ_k is called x**'s semi-invariants or cumulants.**

Show that (neglecting questions of existence and convergence)

$$\begin{aligned} \kappa_1 &= \mu_1 & \mu_1 &= \kappa_1 \\ \kappa_2 &= \mu_2 - \mu_1^2 & \mu_2 &= \kappa_2 + \kappa_1^2 \\ \kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 & \mu_3 &= \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3 \\ &\cdots & &\cdots \end{aligned} \tag{24}$$

***Exercise 9.** Show that

$$\gamma_x(1) = 1, \quad \mathfrak{M}\{x\} = \left[\frac{\partial\gamma_x(t)}{\partial t}\right]_{t=1}, \quad \mathfrak{M}\{x^2\} = \left[\frac{\partial^2\gamma_x(t)}{\partial t^2} + \frac{\partial\gamma_x(t)}{\partial t}\right]_{t=1} \tag{25}$$

$$\mu_x(0) = 1, \quad \mathfrak{M}\{x\} = \left[\frac{\partial\mu_x(t)}{\partial t}\right]_{t=0}, \quad \mathfrak{M}\{x^2\} = \left[\frac{\partial^2\mu_x(t)}{\partial t^2}\right]_{t=0} \tag{26}$$

$$\chi_x(0) = 1, \quad \mathfrak{M}\{x\} = -i\left[\frac{\partial\chi_x(t)}{\partial t}\right]_{t=0}, \quad \mathfrak{M}\{x^2\} = -\left[\frac{\partial^2\chi_x(t)}{\partial t^2}\right]_{t=0}. \tag{27}$$

Check these formulae for the distributions mentioned in Exercise 7.

***Exercise 10.** Often a probability problem may be more easily solved by first finding either the generating, the moment generating, or the characteristic function. Deduce from (4.3.6) a differential equation for the generating function $\gamma(u)$, solve it, and obtain (4.3.7) by expanding the solution into a power series in u.

§ 5.3. As stressed previously, a distribution is not uniquely determined, but only roughly characterized, by any measure of location. However, for practical purposes, it is often sufficient to know the value of such a measure, e.g., the mean, together with a **measure of dispersion,** i.e., of how much the probability mass is spread about the chosen measure of location (although, of course, $\Phi(t)$ is not uniquely determined by these two numbers alone). Here also we may choose this measure in infinitely many ways, and again preference is only a question of convenience. In practice the following measures are used: the **mean deviation,** i.e., $\mathfrak{M}\{|x - \mu|\}$, μ being the mean value; the **dispersion** $\sigma\{x\}$, i.e., in the mechanical picture of § 4.6 the square root of the moment of inertia of the total probability mass with respect to $x = \mu$; one half the **half-width,** γ, i.e., $\gamma = \frac{1}{2}(t_2 - t_1)$, in which t_1, t_2 are the two roots of $\varphi(t) = \frac{1}{2}\varphi_{\max}$, $\varphi_{\max}$ being the maximum value of $\varphi(t)$ assumed to allow only two roots in this equation.

Other measures used are the **semi-interquartile range,** i.e., $\frac{1}{2}(t_{3/4} - t_{1/4})$, in which $t_{3/4}$, $t_{1/4}$, satisfying $\Phi(t) = \frac{3}{4}$, $\Phi(t) = \frac{1}{4}$, respectively, are called the **quartiles;** or more generally any **semi-interpercentile range,** i.e., $\frac{1}{2}(t_{1-\theta} - t_\theta)$, in which t_θ, satisfying $\Phi(t) = \theta < 1$, is called the **percentile** or **fractile** corresponding to the fraction θ (if $\theta = p/10$, $p = 1, 2, \cdots, 9$, t_θ is also called a **decile**); obvi-

ously these concepts are generalizations of the median, which is exactly $t_{1/2}$. (We note that for discontinuous distributions t_0 need not be uniquely defined.) If x is bounded (cf. Example 1, §5.1,) one-half the **range,** i.e., $\frac{1}{2}(G - g)$, is sometimes a convenient measure of dispersion.

Since the rules of calculation are especially simple for the dispersion $\sigma\{x\}$, this measure of dispersion is used mostly, although others are also used, e.g., the half-width is frequently used in physics (cf. Example 10). From the definition of the moment of inertia of a body $\sigma\{x\}$, or σ_x, or briefly σ, is defined by[1]

$$\sigma^2\{x\} = \begin{cases} \sum_{i=-\infty}^{\infty} (t_i - \mu)^2 \varphi_i & \text{for discontinuous distributions} \\ \int_{-\infty}^{\infty} (t - \mu)^2 \varphi(t)\, dt & \text{for continuous distributions} \end{cases} \tag{1}$$

(t_i, φ_i are defined in §4.3, $\varphi(t)$ in §4.4, and μ is the mean). $\delta = \sigma/\mu$ is called the **relative dispersion** or **coefficient of variation.**

Example 1. For a continuous distribution (1) shows that, owing to the piecewise continuity of $\varphi(t) \geqq 0$, $\sigma^2 > 0$. On the other hand, for a discontinuous distribution we may have $\sigma^2 = 0$, but then (1) shows that the distribution is the causal one, $\Phi(t) = \epsilon(t - \mu)$, (4.3.3).

We see from (5.2.1) that σ^2 may also be written

$$\sigma^2\{x\} = \mathfrak{M}\{(x - \mu)^2\}. \tag{2}$$

(Thus σ^2 is x's central moment of second order, cf. Example 2, §5.2.)

Exercise 1. Show that if a is a constant

$$\sigma^2\{ax\} = a^2\sigma^2\{x\} \tag{3}$$

$$\sigma^2\{x + a\} = \sigma^2\{x\}. \tag{4}$$

Exercise 2. By the **relative deviation** of x we understand $y = (x - \mu)/\sigma$. Show that $\mathfrak{M}\{y\} = 0$ and $\sigma\{y\} = 1$. y is said to be a **normalized** (or **standardized**) random variable. We see that to normalize x simply means to choose the zero point and the scale on the x-axis in an especially convenient way.

Exercise 3. If a normalized variable y has the distribution function $\Phi(t)$ and the probability density $\varphi(t)$, show that $x = \sigma y + \mu$ has the distribution function $\Phi\left(\frac{t - \mu}{\sigma}\right)$, the probability density $\frac{1}{\sigma}\varphi\left(\frac{t - \mu}{\sigma}\right)$, $\mathfrak{M}\{x\} = \mu$, and $\sigma\{x\} = \sigma$.

[1] σ^2 itself is called the **variance** of x, $\mathfrak{V}\{x\}$. σ is also called **standard deviation, root-mean-square deviation** (or simply mean deviation, not to be confused with that defined above), or **fluctuation.** Also it is often denoted by $D\{x\}$; for the use of { } instead of ( ), cf.[1] p 55. French, **dispersion, écart quadratique moyen,** German, **Streuung, mittlere Abweichung.**

*Exercise 4. If a normalized variable y has the characteristic function $\chi_y(t)$, show that $x = \sigma y + \mu$ has the characteristic function $\chi_x(t) = e^{i\mu t} \cdot \chi_y(\sigma t)$.

For an arbitrary constant, a, we have the identity

$$(x - a)^2 = ((x - \mu) + (\mu - a))^2 = (x - \mu)^2 + 2(x - \mu)(\mu - a) + (\mu - a)^2.$$

Taking the mean of both sides we get

$$\mathfrak{M}\{(x - a)^2\} = \sigma^2\{x\} + (\mu - a)^2. \tag{5}$$

Exercise 5. Verify this. What is the corresponding theorem on moments of inertia (Steiner's theorem)?

If, especially, we put $a = 0$ in (5) we get the important relation, often used for the calculation of σ^2:

$$\sigma^2\{x\} = \mathfrak{M}\{x^2\} - \mathfrak{M}^2\{x\}. \tag{6}$$

Example 2. From Example 2, § 5.2, we see that this may also be written $\sigma^2 = \mu_2 - \mu_1{}^2$ or, from Exercise 6, § 5.2, $\sigma^2 = \kappa_2$.

Exercise 6. Show that we also have

$$\sigma^2\{x\} = \mathfrak{M}\{x(x - 1)\} - \mu(\mu - 1), \tag{7}$$

which formula is often more convenient than (6), especially for discontinuous distributions.

*Example 3. From Example 2, § 5.2, we see that (7) may also be written $\sigma^2 = \mu_{(2)} - \mu_1(\mu_1 - 1)$.

Example 4. For the variable x in Example 2, § 5.1, we have from (5.1.3)

$$\sigma^2\{x\} = (1 - \theta)^2\theta + (0 - \theta)^2(1 - \theta) = \theta(1 - \theta). \tag{8}$$

Example 5. For the variable x in Example 3, § 5.1, we have from a well-known algebraic theorem

$$\mathfrak{M}\{x^2\} = \sum_{i=1}^{\nu} i^2 \frac{1}{\nu} = \frac{\nu(\nu + 1)(2\nu + 1)}{6} \frac{1}{\nu} = \frac{(\nu + 1)(2\nu + 1)}{6},$$

i.e., from (6) and (5.1.4)

$$\sigma^2\{x\} = \frac{(\nu + 1)(2\nu + 1)}{6} - \left(\frac{\nu + 1}{2}\right)^2 = \frac{\nu^2 - 1}{12}. \tag{9}$$

Example 6. For the binomial distribution (Example 2, § 4.3) we have, from the binomial theorem,

$$\mathfrak{M}\{x(x-1)\} = \sum_{i=0}^{\nu} i(i-1)\binom{\nu}{i}\theta^i(1-\theta)^{\nu-i} =$$

$$\nu(\nu-1)\theta^2 \cdot \sum_{i=2}^{\nu}\binom{\nu-2}{i-2}\theta^{i-2}(1-\theta)^{(\nu-2)-(i-2)} = \nu(\nu-1)\theta^2,$$

i.e., from (7) and (5.1.5)

$$\sigma^2\{x\} = \nu(\nu-1)\theta^2 - (\nu\theta)(\nu\theta-1) = \nu\theta(1-\theta). \tag{10}$$

Later we shall give a much simpler proof of this formula (cf. Example 1, § 6.4).

Example 7. For Poisson's distribution (Example 3, § 4.3) we find

$$\mathfrak{M}\{x(x-1)\} = \sum_{i=0}^{\infty} i(i-1)e^{-\mu}\frac{\mu^i}{i!} = e^{-\mu}\mu^2\sum_{i=2}^{\infty}\frac{\mu^{i-2}}{(i-2)!} = \mu^2,$$

i.e., from (7) and (5.1.6)

$$\sigma^2\{x\} = \mu^2 - \mu(\mu-1) = \mu. \tag{11}$$

Thus the relative dispersion is $\delta = \sigma/\mu = 1/\sqrt{\mu}$, i.e., δ decreases for increasing μ, which means that the probability mass becomes more and more concentrated about μ.

Example 8. For Pascal's distribution (Example 6, § 4.3) we find by the same device, putting $\mu/(1+\mu) = x$, as was used in calculating $\mathfrak{M}\{x\}$ (Example 7, § 5.1)

$$\mathfrak{M}\{x(x-1)\} = \sum_{i=0}^{\infty} i(i-1)\frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^i =$$

$$\frac{x^2}{1+\mu}\frac{d^2}{dx^2}\sum_{i=0}^{\infty}x^i = \frac{x^2}{1+\mu}\frac{d^2}{dx^2}\frac{1}{1-x} = 2\mu^2,$$

i.e., from (7) and (5.1.8)

$$\sigma^2\{x\} = 2\mu^2 - \mu(\mu-1) = \mu^2 + \mu. \tag{12}$$

Example 9. We shall later see (§ 7.3) that for the normal distribution (Example 3, § 4.4) the parameter σ is just the dispersion.

Example 10. For Cauchy's distribution (Example 4, § 4.4) we have

$$\sigma^2\{x\} = \frac{1}{\pi\alpha}\int_{-\infty}^{\infty}(t-\mu)^2\,\frac{dt}{1+\frac{(t-\mu)^2}{\alpha^2}} = \frac{\alpha^2}{\pi}\int_{-\infty}^{\infty}\frac{x^2\,dx}{1+x^2} =$$

$$\frac{\alpha^2}{\pi}\Big|x - \operatorname{arc\,tan} x\Big|_{-\infty}^{\infty} = 2\infty - \alpha^2 = \infty. \quad (13)$$

Thus in this case the dispersion does not exist. On the other hand, one-half the half-width is easily calculated and found to be equal to α, so that this is a convenient measure of dispersion for this distribution which is often met in physics.

Exercise 7. Show by means of the binomial theorem that for Pólya's distribution (Example 7, § 4.3)

$$\mathfrak{M}\{x(x-1)\} = (1+\beta)\mu^2,$$

i.e., from (7) and Exercise 2, § 5.1,

$$\sigma^2\{x\} = \mu(1+\beta\mu), \quad (14)$$

which contains (11) and (12) as special cases, as it should (why?).

Exercise 8. Show by means of Appendix 1 and Exercise 2, § 5.1, that for Laplace's distribution (Example 5, § 4.4)

$$\sigma^2\{x\} = 2\alpha^2. \quad (15)$$

Exercise 9. Show by means of Appendix 1 and (5.1.10) that for the distribution of Example 2, § 4.4,

$$\sigma^2\{x\} = \frac{1}{\lambda^2}. \quad (16)$$

*§ 5.4. By means of Stieltjes' integral, defined in § 4.7, we define for an arbitrary distribution

$$\mu = \mathfrak{M}\{x\} = \int_{-\infty}^{\infty} t\,d\Phi(t). \quad (1)$$

For an arbitrary function $y = f(x)$ we then have

$$\mathfrak{M}\{y\} = \mathfrak{M}\{f(x)\} = \int_{-\infty}^{\infty} f(t)\,d\Phi(t). \quad (2)$$

Putting $y = (x-\mu)^2$ we get

$$\sigma^2\{x\} = \int_{-\infty}^{\infty}(t-\mu)^2\,d\Phi(t). \quad (3)$$

For $y = x^k$ we get

$$\mathfrak{M}\{x^k\} = \int_{-\infty}^{\infty} t^k\,d\Phi(t) = \mu_k. \quad (4)$$

For $y = e^{itx}$ we get

$$\mathfrak{M}\{e^{itx}\} = \int_{-\infty}^{\infty} e^{itx}\, d\Phi(x) = \chi_x(t), \tag{5}$$

and so forth. We leave it to the reader to check that these formulae reduce to the previous corresponding formulae for discontinuous and continuous distributions respectively.

§ 5.5. For a two-dimensional random variable $z = (x, y)$ we again define the mean value, if it exists, as the center of gravity of the whole probability mass, i.e., by its "static moments" with respect to the t- and u-axes:

$$\mathfrak{M}\{z\} = (\mu_x, \mu_y) = \begin{cases} \left(\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} t_i\varphi_{ij}, \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} u_j\varphi_{ij}\right) = \\ \left(\sum_{i=-\infty}^{\infty} t_i\varphi_{i\cdot}, \sum_{j=-\infty}^{\infty} u_j\varphi_{\cdot j}\right) \text{ for discontinuous distributions,} \\ \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t\varphi(t,u)\,dt\,du, \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u\varphi(t,u)\,dt\,du\right) = \\ \left(\int_{-\infty}^{\infty} t\varphi_x(t)\,dt, \int_{-\infty}^{\infty} u\varphi_y(u)\,du\right) \text{ for continuous distributions.} \end{cases} \tag{1}$$

(For the definition of the letters, cf. § 4.10–§ 4.13; in general, the convergence is assumed to be absolute.) Thus to calculate $\mathfrak{M}\{z\}$ we need to know only the two marginal distributions, since from (1)

$$\mathfrak{M}\{z\} = (\mu_x, \mu_y) = (\mathfrak{M}\{x\}, \mathfrak{M}\{y\}). \tag{2}$$

Example 1. For the two-dimensional normal distribution (4.12.7) we see directly from the two marginal distributions calculated in (4.13.6) and (4.13.7) that the parameters μ_x and μ_y are just the two components of $\mathfrak{M}\{z\}$.

***Example 2.** If we form the means of the conditional distributions (for simplicity we consider only the continuous case, § 4.13), we obtain two functions which are called the **regression** of x on y and of y on x, respectively, provided that they exist:

$$\begin{aligned} \mathfrak{M}\{x|u\} &= \int_{-\infty}^{\infty} t\varphi(t|u)\,dt = \mu_x(u) \\ \mathfrak{M}\{y|t\} &= \int_{-\infty}^{\infty} u\varphi(u|t)\,du = \mu_y(t). \end{aligned} \tag{3}$$

We note that in general these two functions, $t = \mu_x(u)$ and $u = \mu_y(t)$, are *not* mutually inverse functions. The graphs of the two functions

are called the **regression curves.** For a normal distribution, especially, we see from (4.13.13) and (4.13.14) that the regression curves are straight lines intersecting in (μ_x, μ_y):

$$\mu_x(u) = \mu_x + \rho \frac{\sigma_x}{\sigma_y}(u - \mu_y)$$
$$\mu_y(t) = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(t - \mu_x). \tag{4}$$

Also for non-normal distributions these straight lines exist (cf. § 5.6) and are called the **mean-square regression lines,** and the coefficients of u and t are called the **regression coefficients.** We note that the normal regression curves coincide if and only if $\rho = \pm 1$. However, in that case the whole probability mass is concentrated along the regression line; i.e., x and y are proportional, and therefore the distribution reduces to a one-dimensional distribution (cf. Example 1, § 4.12). The concept of regression plays an important role in statistics. For further discussion we must, however, refer to a textbook in statistics, e.g., that of Cramér.

For a function $w = f(z) = f(x, y)$ of the random variable z, where w may be either one- or two-dimensional, we define as a generalization of (5.2.1) the mean of w, if it exists, by

$$\mathfrak{M}\{w\} = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} f(t_i, u_j)\varphi_{ij} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t, u)\varphi(t, u)\, dt\, du. \end{cases} \tag{5}$$

For it can again be shown that this is the same result as we would find by first working out the distribution of w (e.g., as discussed in § 4.14) and then applying (5.3.1).[1] We stress that in general we have, cf. (5.2.6),

$$\mathfrak{M}\{f(z)\} \neq f(\mathfrak{M}\{z\}). \tag{6}$$

In the next chapter we shall investigate the two simple functions $w = x + y$ and $w = x \cdot y$.

***Example 3.** The quantities defined in Example 2, § 5.2, may easily be generalized to two-dimensional distributions. Thus, for continuous distributions, e.g., the moments of order k (with respect to the point (0, 0)) are defined, if they exist, by

$$\mathfrak{M}\{x^l y^m\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^l y^m \varphi(x, y)\, dx\, dy, \qquad l + m = k. \tag{7}$$

[1] See, e.g., Cramér, *Random Variables*, p. 19.

Next the characteristic function is defined by

$$\mathfrak{M}\{\exp[i(tx + uy)]\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp[i(tx + uy)]\varphi(x, y)\,dx\,dy =$$

$$\chi_{x,y}(t, u), \quad (8)$$

which function of t and u is seen to exist for all values of t, u and all distributions. If it may be expanded into a power series of it and iu, the coefficient of $(it)^l(iu)^m$ is seen to be the corresponding moment $\mathfrak{M}\{x^l y^m\}$. Also in the two-dimensional case the distribution function, Φ, is uniquely determined by the characteristic function, Φ being obtained from $\chi(t, u)$ by the two-dimensional generalizations of (5.2.14) and (5.2.15).

*Exercise 1. Show that the characteristic functions of the two marginal distributions is simply $\chi_x(t) = \chi_{x,y}(t, 0)$ and $\chi_y(u) = \chi_{x,y}(0, u)$, respectively.

*Exercise 2. Show that the characteristic function of the two-dimensional normal distribution is the simple generalization of (5.2.21)

$$\chi_{x,y}(t, u) = \exp[i(\mu_x t + \mu_y u) - \tfrac{1}{2}(\sigma_x^2 t^2 + 2\rho\sigma_x\sigma_y tu + \sigma_y^2 u^2)]. \quad (9)$$

§ 5.6. In the mechanical picture of § 4.6 we may also speak of the four "quadratic moments" of the probability mass with respect to (μ_x, μ_y), thereby obtaining a natural generalization of the dispersion of a one-dimensional distribution:[1]

$$\begin{aligned} \mathfrak{M}\{(x - \mu_x)^2\} &= \mu_{xx} = \sigma_x^2 \\ \mathfrak{M}\{(y - \mu_y)^2\} &= \mu_{yy} = \sigma_y^2 \\ \mathfrak{M}\{(x - \mu_x)(y - \mu_y)\} &= \mu_{xy} = \mu_{yx}. \end{aligned} \quad (1)$$

Next we introduce the **correlation coefficient** $\rho\{x, y\}$, ρ_{xy}, or briefly ρ, between x and y by the definition

$$\rho\{x, y\} = \rho = \frac{\mu_{xy}}{\sigma_x\sigma_y}. \quad (2)$$

As shown in the example, § 6.1,

$$-1 \leqq \rho \leqq 1. \quad (3)$$

Exercise. Show that if x and y are proportional we have equality in (3), i.e.,

$$\begin{aligned} \rho &= 1 \quad \text{for} \quad x = a^2 y \\ \rho &= -1 \quad \text{for} \quad x = -a^2 y, \end{aligned} \quad (4)$$

where a is an arbitrary real number ($\neq 0$).

[1] σ_x^2 and σ_y^2 are called the **variances**, $\mathfrak{V}\{x\}$ and $\mathfrak{V}\{y\}$, and $\mu_{xy} = \mu_{yx}$ the **covariances** $\mathfrak{CV}\{x, y\}$.

In general, $|\rho| < 1$. In §6.2 we show that if x and y are independent, $\rho = 0$. Thus ρ may be taken as a measure of the dependency, or correlation, between x and y, hence the name. However, the converse is not true, i.e., if $\rho\{x, y\} = 0$, x and y need *not* be independent (cf. Example 1, §6.2). If $\rho = 0$, x and y are sometimes called **uncorrelated.**

Example. For the two-dimensional normal distribution (4.12.7), the expressions (4.13.6) and (4.13.7) for the marginal distributions show immediately that the parameters σ_x and σ_y are just the quadratic moments defined in (1). Later we shall see (§7.6) that the parameter ρ is also precisely the correlation coefficient defined in (2).

We leave it to the reader to generalize the content of §5.5–§5.6 from two- to many-dimensional distributions.

6.

MEAN VALUE AND DISPERSION OF SUMS, PRODUCTS, AND OTHER FUNCTIONS

§6.1. If in (5.5.5) we put $\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{y}$, we get (assuming the convergencies to be absolute)

$$\mathfrak{M}\{\boldsymbol{x}+\boldsymbol{y}\} = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}(t_i+u_j)\varphi_{ij} = \sum_{i=-\infty}^{\infty} t_i\varphi_{i\cdot} + \sum_{j=-\infty}^{\infty} u_j\varphi_{\cdot j} =$$

$$\mathfrak{M}\{\boldsymbol{x}\} + \mathfrak{M}\{\boldsymbol{y}\} \quad (1)$$

and

$$\mathfrak{M}\{\boldsymbol{x}+\boldsymbol{y}\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t+u)\varphi(t,u)\,dt\,du =$$

$$\int_{-\infty}^{\infty} t\varphi_x(t)\,dt + \int_{-\infty}^{\infty} u\varphi_y(u)\,du = \mathfrak{M}\{\boldsymbol{x}\} + \mathfrak{M}\{\boldsymbol{y}\} \quad (2)$$

for discontinuous and continuous distributions, respectively.

The mean value of the sum of two, not necessarily independent, random variables $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *is equal to the sum of the respective mean values*

$$\mathfrak{M}\{\boldsymbol{x}+\boldsymbol{y}\} = \mathfrak{M}\{\boldsymbol{x}\} + \mathfrak{M}\{\boldsymbol{y}\}. \quad (3)$$

Exercise. Show that μ_{xy} defined in (5.6.2) may also be written

$$\mu_{xy} = \mu_{yx} = \mathfrak{M}\{\boldsymbol{xy}\} - \mathfrak{M}\{\boldsymbol{x}\}\mathfrak{M}\{\boldsymbol{y}\}. \quad (4)$$

Example. We can now prove (5.6.3). Let a and b be two arbitrary real numbers, and let us consider the function

$$F = ((\boldsymbol{x}-\mu_x)a + (\boldsymbol{y}-\mu_y)b)^2 \geqq 0.$$

Forming the mean value on both sides we get

$$\mathfrak{M}\{F\} = \mathfrak{M}\{(\boldsymbol{x}-\mu_x)^2\}a^2 + 2\mathfrak{M}\{(\boldsymbol{x}-\mu_x)(\boldsymbol{y}-\mu_y)\}ab$$

$$+ \mathfrak{M}\{(\boldsymbol{y}-\mu_y)^2\}b^2 = \mu_{xx}a^2 + 2\mu_{xy}ab + \mu_{yy}b^2 \geqq 0.$$

The condition that this homogeneous quadratic form in a and b be non-negative is then

$$\mu_{xx}\mu_{yy} - \mu_{xy}^2 \geqq 0, \quad \text{i.e.,} \quad \frac{\mu_{xy}^2}{\mu_{xx}\mu_{yy}} = \rho^2 \leqq 1, \quad \text{Q.E.D.} \quad (5)$$

§ 6.2. If in (5.5.5) we put $w = x \cdot y$, we get, *now assuming* x *and* y *to be independent* (and the convergencies to be absolute),

$$\mathfrak{M}\{xy\} = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} t_i u_j \varphi_{ij} = \sum_{i=-\infty}^{\infty} t_i \varphi_{i\cdot} \sum_{j=-\infty}^{\infty} u_j \varphi_{\cdot j} = \mathfrak{M}\{x\}\mathfrak{M}\{y\} \tag{1}$$

and

$$\mathfrak{M}\{xy\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} tu\varphi(t, u)\, dt\, du = \int_{-\infty}^{\infty} t\varphi_x(t)\, dt \int_{-\infty}^{\infty} u\varphi_y(u)\, du = \mathfrak{M}\{x\}\mathfrak{M}\{y\} \tag{2}$$

for discontinuous and continuous distributions, respectively.

The mean value of the product of two **independent** *random variables* x *and* y *is equal to the product of the respective mean values*

$$\mathfrak{M}\{xy\} = \mathfrak{M}\{x\}\mathfrak{M}\{y\}. \tag{3}$$

We note that (6.1.3) and (3) have no analogues for the mode or the median. This is the reason why the mean is as a rule preferred as a measure of location. From (6.1.4) we see that the necessary and sufficient condition for (3) is that x and y are uncorrelated, i.e., $\rho\{x, y\} = 0$. Thus, if x and y are independent, they are also uncorrelated, but in general the converse is not true, i.e., from $\rho\{x, y\} = 0$ we *cannot* conclude that x and y are independent.

Example 1. For the two-dimensional distribution given in (4.12.8) we get (from symmetry $\mu_x = \mu_y = 0$)

$$\mu_{xy} = \frac{1}{\pi} \int_{-\infty}^{\infty} t\, dt \int_{-\infty}^{\infty} \frac{u\, du}{(1 + t^2 + u^2)^2} = 0 \tag{4}$$

because the integrand in the u-integration is odd. Since σ_x and σ_y both exist (the respective integrals being convergent), ρ exists and is $= 0$. However, x and y are obviously *not* independent in this case since $\varphi(t, u)$ cannot be written as $\varphi_x(t)\varphi_y(u)$ (cf. § 4.13).

In special cases $\rho = 0$ may imply that x and y are independent, e.g., if $z = (x, y)$ is normally distributed (cf. the example, § 12.7). For, in this case, (4.12.7) shows that for $\rho = 0$, $\varphi(t, u) = \varphi_x(t)\varphi_y(u)$, i.e., x and y are independent. However, because in practice it is often reasonable to assume (x, y) to be normally distributed, or because we merely want to make use of (3), we may often treat x and y as being independent if $\rho = 0$ only. *Thus the correlation coefficient plays an important role in many statistical investigations* (cf. § 9.5 and § 11.13).

Exercise 1. For arbitrary constants a_1, a_2, $b_1(>0)$, and $b_2(>0)$, show that

$$\mu_{xy} = \mathfrak{M}\{xy\} - \mathfrak{M}\{x\}\mathfrak{M}\{y\} = \mathfrak{M}\{(x-a_1)(y-a_2)\} - \mathfrak{M}\{x-a_1\}\mathfrak{M}\{y-a_2\} \tag{5}$$

and

$$\rho\left\{\frac{x-a_1}{b_1}, \frac{y-a_2}{b_2}\right\} = \rho\{x, y\}. \tag{6}$$

***Example 2.** From (3) follows an important property of the characteristic functions (Example 2, § 5.2) which indicates their great value. Since, for two *independent* random variables x and y

$$\chi_{x+y}(t) = \mathfrak{M}\{e^{i(x+y)t}\} = \mathfrak{M}\{e^{ixt}\cdot e^{iyt}\} = \mathfrak{M}\{e^{ixt}\}\mathfrak{M}\{e^{iyt}\} = \chi_x(t)\chi_y(t), \tag{7}$$

the characteristic function of a *sum* of independent random variables is equal to the *product* of the respective characteristic functions. Comparing this with (4.14.6) we see that it is much simpler to form the characteristic function of a sum than to form its distribution function. (The same law obviously also holds for the generating and the moment generating functions.) We note that the logarithm of the characteristic function of a *sum* is the *sum* of the logarithms of the respective characteristic functions. This also holds true for each of the coefficients in the power-series expansion of a characteristic function—hence the name "semi-invariants" or "cumulants" for these coefficients (cf. Exercise 8, § 5.2).

***Exercise 2.** Let $x_1, \cdots, x_\nu$ be ν independent random variables with the characteristic functions $\chi_1(t), \cdots, \chi_\nu(t)$. Then show that $z = a_1x_1 + \cdots + a_\nu x_\nu$, where $a_1, \cdots, a_\nu$ are arbitrary constants, has the characteristic function (cf. Exercise 6, § 5.2)

$$\chi_z(t) = \chi_1(a_1t)\cdots\chi_\nu(a_\nu t). \tag{8}$$

***Exercise 3.** For the product of two independent random variables we have no such simple expression for the characteristic function as for the sum. However, in many physical problems we meet multiplicative processes such as the fission chains in atomic bombs, electron avalanches in Geiger-Müller counters, and cascade showers in cosmic rays. Here one primary "particle" (e.g., a neutron) causes a process (e.g., a fission chain) in which the primary particle disappears, but which itself results in the creation of a certain number of secondary "particles" (e.g., the neutrons produced in the chain reaction). Let x, y, and z be the random variable which gives the number of primary particles, of secondary particles created by one primary particle, and the total number of particles created, respectively, and let $\gamma_x(t)$, $\gamma_y(t)$, and $\gamma_z(t)$ be the corresponding generating functions. Then show that

$$\gamma_z(t) = \gamma_x(\gamma_y(t)). \tag{9}$$

By means of (5.2.25) next show that

$$\mathfrak{M}\{z\} = \mathfrak{M}\{x\}\mathfrak{M}\{y\} \tag{10}$$

$$\sigma^2\{z\} = \sigma^2\{x\}\mathfrak{M}^2\{y\} + \sigma^2\{y\}\mathfrak{M}\{x\}. \tag{11}$$

§6.3. From the foregoing results we find for the dispersion of two random variables x and y

$$\sigma^2\{x \pm y\} = \sigma^2\{x\} + \sigma^2\{y\} \pm 2\rho\{x, y\}\sigma\{x\}\sigma\{y\}. \tag{1}$$

This we shall call the **variance law.**

Exercise 1. Verify this. Find the corresponding formula for $\sigma^2\{ax + by\}$, a and b being constants.

In particular, if x and y are independent, or only uncorrelated, (1) becomes

$$\sigma^2\{x \pm y\} = \sigma^2\{x\} + \sigma^2\{y\}. \tag{2}$$

For two uncorrelated random variables the square of the dispersion of the sum is equal to the sum of the squares of the two separate dispersions. We note that the right-hand side of (2) is the same whether we have $+$ or $-$ on the left-hand side. Thus, in general, the relative dispersion of the difference is much larger than that of the sum.

Exercise 2. Let us consider a throw with two dice, and let x and y denote the result on the first and second die respectively. Then from (5.1.4) $\mathfrak{M}\{x\} = \mathfrak{M}\{y\} = 7/2$ and from (5.3.9) $\sigma^2\{x\} = \sigma^2\{y\} = 35/12$. Find the possible values of $x + y$ and $x \cdot y$ and their probabilities, and from this result calculate $\mathfrak{M}\{x + y\}$, $\mathfrak{M}\{x \cdot y\}$, and $\sigma^2\{x + y\}$. Check that these results agree with (6.1.3), (6.2.3), and (6.3.2).

§6.4. The foregoing formulae may immediately be generalized to sums of more than two random variables. Let $x_1, x_2, \cdots, x_\nu$ be ν arbitrary random variables with means $\mu_1, \mu_2, \ldots, \mu_\nu$ and dispersions $\sigma_1, \sigma_2, \ldots, \sigma_\nu$, and let $a_1, a_2, \ldots, a_\nu$ be ν arbitrary constants. For the random variable

$$z = a_1x_1 + a_2x_2 + \cdots + a_\nu x_\nu \tag{1}$$

we then have

$$\mu_z = \mathfrak{M}\{z\} = a_1\mu_1 + a_2\mu_2 + \cdots + a_\nu\mu_\nu \tag{2}$$

and the **general variance law**

$$\begin{aligned}\sigma_z^2 = \sigma^2\{z\} = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_\nu^2\sigma_\nu^2 + \\ 2a_1a_2\rho_{12}\sigma_1\sigma_2 + \cdots + 2a_1a_\nu\rho_{1\nu}\sigma_1\sigma_\nu + \\ \ldots \quad \ldots \quad \ldots + 2a_{\nu-1}a_\nu\rho_{\nu-1,\nu}\sigma_{\nu-1}\sigma_\nu\end{aligned} \tag{3}$$

in which we have put $\rho_{ij} = \rho\{x_i, x_j\}$.

Exercise 1. Verify this.

In particular, if $x_1, \cdots, x_\nu$ are uncorrelated, i.e., two-by-two uncorrelated, $\rho_{ij} = 0$ for $i \neq j$, (3) becomes

$$\sigma_z^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_\nu^2\sigma_\nu^2. \tag{4}$$

We often find cases in which in (1) $\mu_1 \sim \mu_2 \sim \cdots$, $\sigma_1 \sim \sigma_2 \sim \cdots$, and $a_1 \sim a_2 \sim \cdots$. Then $\mathfrak{M}\{z\} \sim \text{const}\ \nu$ and $\sigma^2\{z\} \sim \text{const}\ \nu$. Thus for the relative dispersion of z we have $\sigma/\mu \sim \text{const}/\sqrt{\nu} \underset{\nu\to\infty}{\to} 0$. This is the reason that, e.g., in applications of statistical mechanics (cf. §4.16) the deviations from the means are as a rule negligible.

In (1) putting $a_1 = a_2 = \cdots = a_\nu = 1/\nu$, z becomes the *arithmetic average* of $x_1, \cdots, x_\nu$[1], which we shall denote by $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_\nu}{\nu}. \tag{5}$$

From (2) we then have

$$\mathfrak{M}\{\bar{x}\} = \frac{\mu_1 + \mu_2 + \cdots + \mu_\nu}{\nu} = \overline{\mathfrak{M}\{x\}}. \tag{6}$$

Furthermore, assuming $x_1, \cdots, x_\nu$ to be uncorrelated we have from (4)

$$\sigma^2\{\bar{x}\} = \frac{{\sigma_1}^2 + {\sigma_2}^2 + \cdots + {\sigma_\nu}^2}{\nu^2} = \frac{1}{\nu}\overline{\sigma^2\{x\}}. \tag{7}$$

These formulae are of special interest if we interpret $x_1, x_2, \cdots, x_\nu$ as ν independent observations of one and the same random variable x with mean μ and dispersion σ. Since we then have $\mu_1 = \mu_2 = \cdots = \mu_\nu = \mu$ and $\sigma_1 = \sigma_2 = \cdots = \sigma_\nu = \sigma$, (6) and (7) reduce to

$$\mathfrak{M}\{\bar{x}\} = \mu \tag{8}$$

$$\sigma^2\{\bar{x}\} = \frac{\sigma^2}{\nu}. \tag{9}$$

These formulae are of the greatest importance in the theory of errors and elsewhere (cf. Chapter 11).

Exercise 2. Often we do not form the simple average (5), but the **weighted average**

$$\bar{x}^{(p)} = \frac{p_1 x_1 + \cdots + p_\nu x_\nu}{p_1 + \cdots + p_\nu}, \tag{10}$$

in which $p_1, \cdots, p_\nu$ are arbitrary non-negative constants called the **weights**. Find $\mathfrak{M}\{\bar{x}^{(p)}\}$ and $\sigma^2\{\bar{x}^{(p)}\}$ in the general case and in the special case $\mu_1 = \mu_2 = \cdots = \mu$ and $\sigma_1 = \sigma_2 = \cdots = \sigma$.

Example 1. The above formulae are often more convenient for the evaluation of $\mathfrak{M}\{x\}$ and $\sigma\{x\}$ than the direct evaluation from the defining equations, e.g., in the case of the binomial distribution (cf.

[1] Also often called the *mean*. In order not to confuse this word with its other application, $\mathfrak{M}\{x\}$, we shall always speak of an *average* in formulae like (5) (cf. footnote 1, p. 55).

Examples 4, §5.1, and 6, §5.3). Let all the variables $x_1, \cdots, x_\nu$ be of the type discussed in Example 2, §5.1; i.e., from (5.1.3) $\mu_1 = \cdots = \mu_\nu = \theta$ and from (5.3.8) ${\sigma_1}^2 = \cdots = {\sigma_\nu}^2 = \theta(1-\theta)$. Then the random variable x in Bernoulli's problem (§3.7), the distribution of which is the binomial distribution, may obviously be written

$$x = x_1 + x_2 + \cdots + x_\nu. \tag{11}$$

Thus from (2) we find immediately

$$\mathfrak{M}\{x\} = \nu\theta, \tag{12}$$

in agreement with (5.1.5), and from (4)

$$\sigma^2\{x\} = \nu\theta(1-\theta) \tag{13}$$

in agreement with (5.3.10), but by a much simpler calculation than in the direct evaluation. For the relative frequency $f = x/\nu$ we have

$$\mathfrak{M}\{f\} = \theta \tag{14}$$

and

$$\sigma^2\{f\} = \frac{\theta(1-\theta)}{\nu}. \tag{15}$$

Since $y = \theta(1-\theta)$ in the interval $0 \leqq \theta \leqq 1$ has the maximum value $\frac{1}{4}$ for $\theta = \frac{1}{2}$, we have

$$\sigma^2\{f\} \leqq \frac{1}{4\nu}. \tag{16}$$

*Example 2. Often it is convenient to introduce the **moment matrix** of $x_1, \cdots, x_\nu$, i.e., the quadratic symmetric matrix of order ν (cf. Appendix 2) defined by

$$\underset{\nu\nu}{M} = \{\mu_{rs}\} = \{\mathfrak{M}\{(x_r - \mu_r)(x_s - \mu_s)\}\} = \{\mathfrak{M}\{x_r x_s\} - \mathfrak{M}\{x_r\}\mathfrak{M}\{x_s\}\}, \tag{17}$$

that is,

$$\mu_{ii} = \sigma^2\{x_i\} = {\sigma_i}^2, \qquad \mu_{ij} = \sigma_i\sigma_j\rho_{ij}. \tag{18}$$

We then see that the necessary and sufficient condition for $x_1, \cdots, x_\nu$ being uncorrelated is that M is a diagonal matrix. Sometimes we also introduce the quadratic symmetric **correlation matrix** of $x_1, \cdots, x_\nu$, defined by

$$C = \left\{\frac{\mu_{rs}}{\sigma_r\sigma_s}\right\} = \{\rho_{rs}\}, \tag{19}$$

in which we have put $\rho_{ii} = \mu_{ii}/\sigma_i\sigma_i = 1$. If Σ denotes the diagonal matrix with $\sigma_1, \sigma_2, \cdots, \sigma_\nu$ as its diagonal elements, we have $C =$

$\boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{M} \cdot \boldsymbol{\Sigma}^{-1}$ (check). Thus another necessary and sufficient condition for $x_1, \cdots, x_\nu$ being uncorrelated is that $\boldsymbol{C} = \boldsymbol{E}$.

*Exercise 3. Often it is useful to perform a linear transformation from one set of random variables $x_1, \cdots, x_\nu$, to another set, $y_1, \cdots, y_\lambda$, where λ need not be equal to ν:

$$y_i = f_{i0} + \sum_{j=1}^{\nu} f_{ij} x_j, \qquad i = 1, 2, \cdots, \lambda, \tag{20}$$

f_{0i} and f_{ij} being constants. Show that the quadratic moments of the y's are given from those of the x's by

$$\mu_{rs}^{(y)} = \sum_{i=1}^{\nu} \sum_{j=1}^{\nu} f_{ri} f_{sj} \mu_{ij}^{(x)}, \qquad r, s = 1, 2, \cdots, \lambda. \tag{21}$$

Next, using matrix symbolism (cf. Appendix 2), show that (20) and (21) may be written in the concise form

$$\underset{\lambda 1}{\boldsymbol{Y}} = \underset{\lambda 1}{\boldsymbol{F}_0} + \underset{\lambda \nu}{\boldsymbol{F}} \cdot \underset{\nu 1}{\boldsymbol{X}} \tag{22}$$

and

$$\underset{\lambda\lambda}{\boldsymbol{M}^{(Y)}} = \underset{\lambda\nu}{\boldsymbol{F}} \cdot \underset{\nu\nu}{\boldsymbol{M}^{(X)}} \cdot \underset{\nu\lambda}{\boldsymbol{F}^*}. \tag{23}$$

§6.5. In §6.4 we have deduced simple formulae to calculate $\mathfrak{M}\{z\}$ and $\sigma\{z\}$, if z is a *linear* function of a number of random variables. However, for practical applications, it is also of importance to carry out such calculations when z is *not* a linear function. For example, in physics, most measurements are *indirect*, being given as non-linear functions of other, directly measured quantities, e.g., the volume of a sphere, $V = (\pi/6)\, d^3$, d being the diameter; and the specific resistance of a wire, $\rho = \pi R d^2/4l$, R being the total resistance, l its length, and d its diameter.

In principle we are also able to treat such cases in which $z = f(x_1, \cdots, x_\nu)$ is a non-linear function, because from the distributions of $x_1, \cdots, x_\nu$ that of z is given (cf. §4.14), and thus $\mathfrak{M}\{z\}$ and $\sigma\{z\}$; but in general these quantities will not be simple functions of the means and dispersions of $x_1, \cdots, x_\nu$. However, often the probability mass of the joint distribution will be concentrated in a relatively narrow region about the point $(\mu_1, \cdots, \mu_\nu)$ and z will be such a slowly varying function that within this region it may be treated as a linear function with sufficient approximation.

Let us consider only continuous distributions, and let us first consider the one-dimensional case, $\nu = 1$, i.e., $z = f(x)$. In any case most of the probability mass of x will lie within the interval from $\mu - a\sigma$ to $\mu + a\sigma$ for some suitably chosen value of the constant a (cf. (8.1.4)). Let us assume, as is often the case in practice, that a

is a small integer, say $a \sim 1$. Since the probability of finding values of x outside the interval $|x - \mu| \sim \sigma$ is thus assumed to be small, we need consider only values of x within this region. Expanding $z = f(x)$ into Taylor's series from the point $x = \mu$, we may write

$$z = f(x) = f(\mu) + (x - \mu)f'(\mu) + \frac{(x - \mu)^2}{2} f''(\xi), \tag{1}$$

ξ being some value between μ and x. Now we need consider only values of x for which $|x - \mu| < \sigma$, and, assuming furthermore, that z is a slowly varying function in this region, i.e., that

$$\frac{\sigma^2}{2}|f''(\mu)| \ll \sigma|f'(\mu)| \qquad \text{or} \qquad \sigma|f''(\mu)| \ll |f'(\mu)|, \tag{2}$$

we may neglect the quadratic term in (1) and write

$$z \sim f(\mu) + (x - \mu)f'(\mu). \tag{3}$$

From (3) we obtain

$$\mathfrak{M}\{z\} \sim f(\mathfrak{M}\{x\}) = f(\mu) \tag{4}$$

$$\sigma\{z\} \sim |f'(\mu)|\sigma\{x\}. \tag{5}$$

Exercise 1. From (3) and Exercise 3, §5.3, show that the distribution of z is approximately given by

$$\Phi_z(u) \sim \Phi_x\left(\frac{u - f(\mu)}{f'(\mu)} + \mu\right), \qquad \varphi_z(u) \sim \frac{1}{|f'(\mu)|}\varphi_x\left(\frac{u - f(\mu)}{f'(\mu)} + \mu\right). \tag{6}$$

Exercise 2. Show that for $z = 1/x$ we have

$$\mathfrak{M}\{z\} \sim \frac{1}{\mu}, \quad \sigma\{z\} \sim \frac{\sigma}{\mu^2}, \quad \text{i.e.,} \quad \frac{\sigma_z}{\mu_z} \sim \frac{\sigma}{\mu}, \tag{7}$$

and generalize this result to $z = x^k$, $k = 0, \pm 1, \pm 2, \cdots$.

Let $z = f(x_1, \cdots, x_\nu)$ be such a slowly varying function in that region, within which most of the probability mass is concentrated, that z may be approximated by its tangent plane. We assume the region to be of the order $|x_i - \mu_i| < \sigma_i$ and obtain as a generalization of (3)

$$z \sim f(\mu_1, \cdots, \mu_\nu) + \frac{\partial f}{\partial x_1}(x_1 - \mu_1) + \cdots + \frac{\partial f}{\partial x_\nu}(x_\nu - \mu_\nu), \tag{8}$$

in which the partial derivatives have to be taken at the point $(\mu_1,$

$\cdots, \mu_\nu)$. Assuming $x_1, \cdots, x_\nu$ to be uncorrelated, (8) gives

$$\mathfrak{M}\{z\} \sim f(\mu_1, \cdots, \mu_\nu) \tag{9}$$

$$\sigma^2\{z\} \sim \left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_1{}^2 + \cdots + \left(\frac{\partial f}{\partial x_\nu}\right)^2 \sigma_\nu{}^2. \tag{10}$$

Exercise 3. Write down the expression corresponding to (10) if the assumption of $x_1, \cdots, x_\nu$ being uncorrelated is no longer fulfilled.

7.

THE NORMAL DISTRIBUTION

§ 7.1. The distribution most important in both theory and practice is the **normal distribution**, which from Example 3, § 4.4, is a continuous distribution with the probability density

$$d\Phi = \varphi(t)\,dt = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt$$

$$\left(= \frac{h}{\sqrt{\pi}}\exp\left[-h^2(t-\mu)^2\right]dt\right). \quad (1)$$

In § 7.2 we shall prove that (1) is normalized to one as it should be, i.e., that

$$\frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt = 1. \quad (2)$$

The graph of $\varphi(t)$ is also called **Gauss' error curve**, and the parameter

$$h = \frac{1}{\sqrt{2}\,\sigma}, \quad (3)$$

often used in older literature, is called the **precision measure** (since the larger h is, the more rapidly the graph of $\varphi(t)$ falls off to 0). Since $\varphi(\mu + t) = \varphi(\mu - t)$, $\varphi(t)$ is symmetric about the line $t = \mu$; and, since

$$\varphi'(t) = \frac{-1}{\sqrt{2\pi}\,\sigma^3}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right](t-\mu) \quad (4)$$

and

$$\varphi''(t) = \frac{1}{\sqrt{2\pi}\,\sigma^3}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]\left(\frac{(t-\mu)^2}{\sigma^2} - 1\right), \quad (5)$$

$\varphi(t)$ has one and only one maximum for $t = \mu$ and two inflection points for $t = \mu \pm \sigma$.

If, especially, $\mu = 0$ and $\sigma = 1$ the distribution is said to be **normalized** (cf. Exercises 1 and 2, § 5.3). In this case the probability den-

sity is denoted by[1]

$$d\Psi = \psi(t)\,dt = \frac{1}{\sqrt{2\pi}}\,e^{-t^2/2}\,dt, \tag{6}$$

the graph of which is shown in Fig. 5 and which is tabulated in Table I.

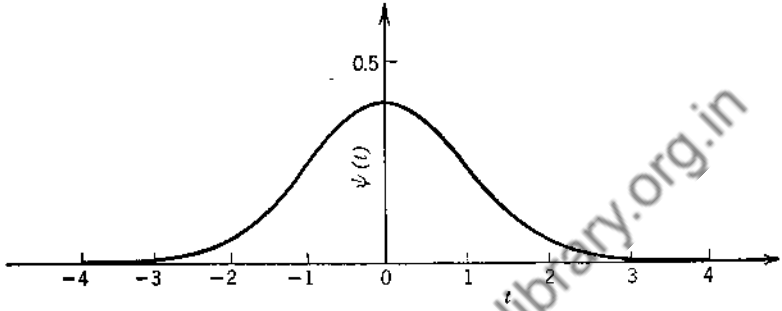

FIG. 5.

For the normal distribution function we have from (1)

$$\Phi(t) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{t}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt, \tag{7}$$

which from (4) has only one inflection point, at $t = \mu$. Since $\varphi(t)$ is symmetric about the line $t = \mu$, we have from (2)

$$\int_{-\infty}^{\mu+t}\varphi(t)\,dt + \int_{\mu+t}^{\infty}\varphi(t)\,dt = \int_{-\infty}^{\mu+t} + \int_{-\infty}^{\mu-t} =$$

$$\Phi(\mu+t) + \Phi(\mu-t) = 1, \tag{8}$$

i.e., in particular,

$$\Phi(\mu) = \tfrac{1}{2}. \tag{9}$$

Thus $\Phi(t)$ is symmetric about the point $(\mu, \frac{1}{2})$.[2] The normalized distribution function, $\mu = 0$ and $\sigma = 1$, is denoted by

$$\Psi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-t^2/2}\,dt. \tag{10}$$

Its graph is shown in Fig. 6, and it is tabulated in Table I. We stress

[1] Instead of ψ the letter φ is often used for the normalized normal distribution. However, this is impossible here since by Φ, φ we denote an arbitrary distribution (for reasons mentioned on p. 9).

[2] We note that this holds for any distribution which is symmetric about $t = \mu$, e.g., also for Cauchy's distribution (4.4.9).

that both $\psi(t)$ and $\Psi(t)$ approach their limiting values very rapidly for $t \to \pm\infty$.

Often $\Psi(t)$ is not tabulated but the so-called **error integral,** or **error function** or **probability integral**[1]

$$\Theta(t) = \operatorname{erf} t = \frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2}\, dt = 2\Psi(\sqrt{2}\, t) - 1. \tag{11}$$

Exercise 1. Verify this.

Exercise 2. If x is normalized and normally distributed, show that $|x|$ has the distribution function $\Theta(t/\sqrt{2})$.

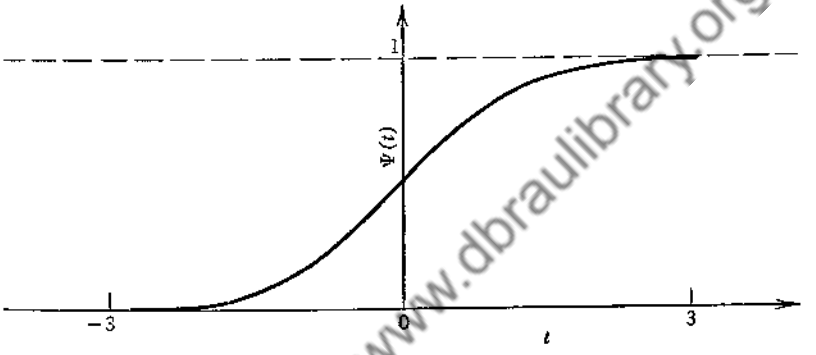

FIG. 6.

Exercise 3. Show that for an arbitrary normal distribution

$$\Phi(t) = \Psi\left(\frac{t-\mu}{\sigma}\right) \qquad \text{and} \qquad \varphi(t) = \frac{1}{\sigma}\psi\left(\frac{t-\mu}{\sigma}\right) \tag{12}$$

(cf. Exercise 3, §5.3). Thus the graphs of $\varphi(t)$ and $\Phi(t)$ are easily obtained from Figs. 5 and 6.

Exercise 4. Show that $y = ax + b$, a ($\neq 0$) and b being arbitrary constants, is also normally distributed.

Exercise 5. If x is normally distributed, show that an arbitrary function $y = f(x)$ is approximately normally distributed with the parameters $\mu_y = f(\mu_x)$ and $\sigma_y = |f'(\mu_x)|\sigma_x$, if f satisfies $\sigma|f''(\mu)| \ll |f'(\mu)|$ (cf. §6.5; expand $f(x)$ by means of Taylor's theorem).

§7.2. We shall now prove (7.1.2). However, the integral cannot be worked out directly, but by using the theory of double and plane

[1] French, **intégrale de Gauss;** German, **Fehlerintegral.** $\theta(t)$ is tabulated to 5 places in Peirce, *A Short Table of Integrals*, Boston, 1929. The most complete tables are found in *Mathematical Tables*, Vol. VII, published by the British Association for the Advancement of Science.

integrals the evaluation can be reduced to the determination of the volume of a body. To that end we form the square of the certainly existing and non-negative number on the left-hand side of (7.1.2). Putting in the first factor $(t - \mu)/\sigma = x$ and in the second $(t - \mu)/\sigma = y$ as new variables, we find

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy \;=$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2} - \frac{y^2}{2} \right] dx\, dy. \quad (1)$$

Thus our task is to find the volume of the body lying between the whole xy-plane and the surface given by $z = \frac{1}{2\pi} \exp\left[-\frac{x^2 + y^2}{2} \right] = \frac{1}{2\pi} \exp\left[-\frac{r^2}{2} \right]$, where r denotes the distance between the points (0, 0) and (x, y). This volume we now find, owing to its symmetry with respect to the z-axis, by dividing it into "infinitely thin" cylinders, the bases of which are rings with radii r and $r + dr$, i.e., area $2\pi r\, dr$, and the heights of which are $\frac{1}{2\pi} \exp\left[-\frac{r^2}{2} \right]$; i.e., the volume of each cylinder is $r \exp\left[-\frac{r^2}{2} \right] dr$. Thus the total volume is, putting $r^2/2 = u$ as a new variable,

$$\int_0^{\infty} r \exp\left[-\frac{r^2}{2} \right] dr = \int_0^{\infty} e^{-u}\, du = 1, \quad (2)$$

which completes our proof.

§ 7.3. For the mean of a normally distributed random variable x we find from (7.1.1)

$$\mathfrak{M}\{x\} = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} t \exp\left[-\frac{(t - \mu)^2}{2\sigma^2} \right] dt =$$

$$\int_{-\infty}^{\infty} ((t - \mu) + \mu)\varphi(t)\, dt = 0 + \mu \cdot 1 = \mu \quad (1)$$

due to (7.1.2) and the symmetry of $\varphi(t)$ about $t = \mu$. For the dispersion we find by partial integration and by putting $(t - \mu)/\sigma = u$ as a new variable

$$\sigma^2\{x\} = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}(t-\mu)^2\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt =$$

$$\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}u\,d\left(-\exp\left[-\frac{u^2}{2}\right]\right) = \sigma^2\left(\left|-\frac{u\exp[-u^2/2]}{\sqrt{2\pi}}\right|_{-\infty}^{\infty} + \right.$$

$$\left.\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\frac{u^2}{2}\right]du\right) = \sigma^2(0+1) = \sigma^2 \quad (2)$$

due to (7.1.2) and the fact that $u\exp[-u^2/2] \to 0$ for $u \to \pm\infty$.

Thus the parameters μ and σ are so chosen as to be equal to the mean and dispersion respectively. This is the reason for writing the seemingly awkward factor 2 in the exponential in (7.1.1).

§7.4. The probability that x assumes a value in the interval between t_1 and t_2 is given by

$$P(t_1 \leqq x \leqq t_2) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{t_1}^{t_2}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt. \quad (1)$$

This formula is of special interest for $t_1 = \mu - a\sigma$ and $t_2 = \mu + a\sigma$, a being an arbitrary positive constant. Putting $(t-\mu)/\sigma = u$ as a new variable, (1) then reduces to

$$P(|x-\mu| \leqq a\sigma) = \frac{1}{\sqrt{2\pi}}\int_{-a}^{a}\exp\left[-\frac{u^2}{2}\right]du = \int_{-\infty}^{a} - \int_{-\infty}^{-a} =$$

$$2\,\Psi(a) - 1 = \Theta\left(\frac{a}{\sqrt{2}}\right), \quad (2)$$

in which we have used (7.1.8) and (7.1.11). This function is tabulated as a function of a in Table I. In particular we find from this table that

$$P(|x-\mu| \leqq \sigma) = 0.6826 \sim \tfrac{2}{3}. \quad (3)$$

The probability that x deviates from μ by less than σ is roughly $\frac{2}{3}$.

Conversely from (2) we can determine the values of a corresponding to given values of P, which we shall call the **tolerance limits**, i.e., the limits within which x lies with a given probability P (assuming this interval to be symmetric with respect to μ). For practical purposes it is, however, more convenient to find a corresponding to given values of the probability of the complementary event, viz., $P(|x-\mu| \geqq a\sigma) = P(a)$. In Table II we give $a = a(P)$.[1] The most important values

[1] This table is reprinted from Table I of Fisher and Yates, *Statistical Tables*, Oliver and Boyd, by permission of the authors and publishers.

are

P	a
0.001	3.29
0.01	2.58
0.05	1.96

(4)

which are called the 0.1%, 1%, and 5% limits, respectively.

Exercise 1. Assume that in shooting with a gun the conditions of Example 1, §4.12, are fulfilled. Find the probability of hitting within a vertical quadratic region with center at $(x, y) = (\mu_x, \mu_y)$, called in ordnance the center of impact, C.I., and sides $2a = 2\sigma_x = 2\sigma_y$.

Exercise 2. Assume the same conditions as in the previous exercise. The distance r from C.I. to the point of impact is then also a random variable. For $\sigma_x = \sigma_y = \sigma$, show that its distribution is given by

$$d\Phi = \frac{1}{\sigma^2} t \exp\left[-\frac{t^2}{2\sigma^2}\right] dt. \tag{5}$$

Next find the probability of hitting within a circle with center at C.I. and radius $R = \sigma$.

Exercise 3. In the exercise, §4.16, we have shown that, for a conservative system consisting of N particles, the three velocity components v_x, v_y, v_z of each particle are mutually independent and that each is normally distributed with the parameters $\mu_x = \mu_y = \mu_z = 0$ and $\sigma_x = \sigma_y = \sigma_z = \sigma$ $(= \sqrt{kT/m})$. Show that the random variable $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$ has the Maxwell-Boltzmann distribution (4.8.2) with $\alpha = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma^3} \left(= \sqrt{\frac{2}{\pi}} \left(\frac{m}{kT}\right)^{3/2} \right)$, $\beta = \frac{1}{2\sigma^2} \left(= \frac{m}{2kT} \right)$.

Example 1. By the **probable deviation** ρ one understands the 50% limit in (2), $a = a(\frac{1}{2})$, i.e., ρ is given by

$$P(|x - \mu| \geqq \rho) = P(|x - \mu| \leqq \rho) = \tfrac{1}{2}. \tag{6}$$

From (2) and Table II this gives

$$\Psi\left(\frac{\rho}{\sigma}\right) = \frac{3}{4}, \qquad \text{i.e.,} \qquad \rho = 0.67449\sigma \sim \frac{2}{3}\sigma. \tag{7}$$

The interval symmetric about μ for which the probabilities of x lying within and without are equal, and equal to $\frac{1}{2}$, is roughly $\frac{2}{3}$.

In older literature ρ is often used as a dispersion parameter in the normal distribution instead of σ, but in newer literature ρ is used less and less frequently,

Example 2. Another dispersion parameter also sometimes used in older literature, but seldom any more, is the **mean deviation** $\theta = \mathfrak{M}\{|x - \mu|\}$. From (7.1.1) we get, introducing $\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2 = u$ as a

new variable

$$\theta = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}|t-\mu|\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt = \frac{2\sigma}{\sqrt{2\pi}}\int_0^{\infty}e^{-u}\,du =$$

$$\sqrt{\frac{2}{\pi}}\,\sigma = 0.79788\sigma \sim \frac{4}{5}\,\sigma. \quad (8)$$

Example 3. Still another dispersion measure, often used in physics, is half the half-width γ. From $\varphi(t) = \frac{1}{2}\varphi(\mu)$ we get $(t-\mu)^2/2\sigma^2 = \ln 2$, i.e.,

$$\gamma = \tfrac{1}{2}(t_2 - t_1) = \sqrt{2\ln 2}\,\sigma = 1.1774\sigma. \quad (9)$$

§7.5. The normal distribution has a number of important properties. We shall state some of them here (cf. also §8.2 and §8.3).

The sum of two independent random variables x *and* y, *which are normally distributed with the parameters* μ_x, σ_x *and* μ_y, σ_y, *respectively, is normally distributed with the parameters* $\mu_{x+y} = \mu_x + \mu_y$ *and* $\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$.

The latter part of this theorem follows from (6.1.3), (6.3.2), and the fact that the parameters represent the mean and the dispersion respectively. The first part follows from (4.14.6). Inserting the normal probability density for φ_x and φ_y this formula gives

$$\varphi_{x+y}(s) = \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty}\exp\left[-\frac{(r-\mu_x)^2}{2\sigma_x^2} - \frac{(s-r-\mu_y)^2}{2\sigma_y^2}\right]dr. \quad (1)$$

Using the expressions for μ_{x+y} and σ_{x+y} the expression in the exponent may be rewritten as

$$\frac{(r-\mu_x)^2}{2\sigma_x^2} + \frac{(s-r-\mu_y)^2}{2\sigma_y^2} =$$

$$\frac{\sigma_{x+y}^2}{2\sigma_x^2\sigma_y^2}\left(r - \frac{\sigma_y^2\mu_x + \sigma_x^2(s-\mu_y)}{\sigma_{x+y}^2}\right)^2 + \left(\frac{s-\mu_{x+y}}{2\sigma_{x+y}^2}\right)^2.$$

Introducing this into (1), putting as a new variable $t = \frac{\sigma_{x+y}}{\sigma_x\sigma_y}\left(r - \frac{\sigma_y^2\mu_x + \sigma_x^2(s-\mu_y)}{\sigma_{x+y}^2}\right)$, and using (7.1.2), we finally find

$$\varphi_{x+y}(s) = \frac{1}{2\pi\sigma_{x+y}}\int_{-\infty}^{\infty}\exp\left[-\frac{t^2}{2} - \frac{(s-\mu_{x+y})^2}{2\sigma_{x+y}^2}\right]dt =$$

$$\frac{1}{\sqrt{2\pi}\,\sigma_{x+y}}\cdot\exp\left[-\frac{(s-\mu_{x+y})^2}{2\sigma_{x+y}^2}\right], \quad (2)$$

which proves our statement.

Example 1. It is noteworthy that this property of the sum having a distribution of the same form as the two addends is not characteristic of the normal distribution alone.[1] For example, we have already proved this for the binomial distribution (Exercise 4, §4.11) and Poisson's distribution (Exercise 5, §4.11). It may also be shown for Cauchy's distribution (cf. problem 41) and for the so-called χ^2-distribution (cf. §7.8, especially Exercise 2).

***Example 2.** From the characteristic functions of the binomial, Poisson's, the normal, and Cauchy's distributions (Exercise 7, §5.2) and the theorem stated in Example 2, §6.2, it follows at once that, if two independent random variables each has the same of these distributions, their sum also has this distribution. This fact among others shows the great utility of the characteristic functions.

Obviously, our theorem may be generalized to the following: *If* $x_1, \cdots, x_\nu$ *are* ν *independent and normally distributed random variables and if* $a_1, \cdots, a_\nu$ *are* ν *arbitrary constants, then*

$$z = a_1 x_1 + \cdots + a_\nu x_\nu \tag{3}$$

is also normally distributed with the parameters given in (6.4.2) *and* (6.4.4).

Exercise 1. Verify this.

Exercise 2. Show the following generalization of Exercise 5, §7.1. If $x_1, \cdots, x_\nu$ are independent and normally distributed with the parameters $\mu_1, \sigma_1, \cdots, \mu_\nu, \sigma_\nu$, respectively, then an arbitrary function $y = f(x_1, \cdots, x_\nu)$ is approximately normally distributed with the parameters

$$\mu_y = f(\mu_1, \cdots, \mu_\nu)$$

and (4)

$$\sigma_y^2 = \left(\frac{\partial f(\mu_1, \cdots, \mu_\nu)}{\partial x_1}\right)^2 \sigma_1^2 + \cdots + \left(\frac{\partial f(\mu_1, \cdots, \mu_\nu)}{\partial x_\nu}\right)^2 \sigma_\nu^2 \tag{5}$$

if f varies only slowly in the region, $|x_i - \mu_i| \sim \sigma_i$, in which the main probability mass of the ν-dimensional random variable $(x_1, \cdots, x_\nu)$ is concentrated (cf. §6.5; expand f by means of Taylor's theorem).

If, especially, in (3) we put $a_1 = a_2 = \cdots = a_\nu = 1/\nu$, i.e., $z = \bar{x}$, and interpret $x_1, \cdots, x_\nu$ as ν independent observations of one and the same random variable x we have the theorem: $\bar{x}$ *is normally distributed with the parameters* μ *and* $\sigma/\sqrt{\nu}$.

Example 3. Without proof we mention that we have the converse theorem that, *if the sum of two independent variables* x *and* y *is normally distributed, then both* x *and* y *themselves are normally distributed.*[2]

[1] Cf. Cramér, *Random Variables*, p. 51.

[2] Cramér, *Random Variables*, p. 52.

§7.6. The two-dimensional normal distribution is given by (4.12.7), i.e.,

$$d\Phi = \varphi(t, u)\, dt\, du =$$

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{1-\rho^2}\left\{\frac{(t-\mu_x)^2}{2\sigma_x{}^2} - \rho\frac{(t-\mu_x)(u-\mu_y)}{\sigma_x\sigma_y} + \frac{(u-\mu_y)^2}{2\sigma_y{}^2}\right\}\right] dt\, du. \quad (1)$$

As shown in the example, §4.13, (1) is normalized to one, i.e.,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(t, u)\, dt\, du = \int_{-\infty}^{\infty} \varphi_x(t)\, dt = \int_{-\infty}^{\infty} \varphi_y(u)\, du = 1. \quad (2)$$

In Example 1, §5.5, and in the example, §5.6, we have shown that the parameters μ_x, μ_y, σ_x and σ_y are so chosen that they denote the means and dispersions, respectively. Finally, it may also be shown that the parameter ρ is so chosen that it denotes the correlation coefficient defined in (5.6.2).

*Example 1. To show this we calculate, putting $x = (t-\mu_x)/\sigma_x$ and $y = (u-\mu_y)/\sigma_y$ as new variables,

$$\mu_{xy} = \mathfrak{M}\{(x-\mu_x)(y-\mu_y)\} =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t-\mu_x)(u-\mu_y)\varphi(t, u)\, dt\, du = \frac{\sigma_x\sigma_y}{2\pi\sqrt{1-\rho^2}}\cdot$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right] dx\, dy. \quad (3)$$

In order to evaluate this rather complicated double integral we introduce two new variables in order to separate them:

$$x = \frac{v-w}{\sqrt{2}}, \quad y = \frac{v+w}{\sqrt{2}}, \quad \text{i.e.,} \quad xy = \frac{v^2-w^2}{2},$$

$$x^2 - 2\rho xy + y^2 = v^2(1-\rho) + w^2(1+\rho). \quad (4)$$

Since in our case the Jacobian (4.14.2) is equal to

$$\frac{\partial(x, y)}{\partial(v, w)} = \begin{vmatrix} \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \end{vmatrix} = \begin{vmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{vmatrix} = 1 \quad (5)$$

and $1 - \rho^2 = (1 - \rho)(1 + \rho)$, we obtain from (4.14.3)

$$\mu_{xy} = \frac{\sigma_x\sigma_y}{4\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (v^2 - w^2)\cdot \exp\left[-\frac{1}{2}\left(\frac{v^2}{1+\rho} + \frac{w^2}{1-\rho}\right)\right] dv\, dw =$$

$$\frac{\sigma_x\sigma_y}{2}\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v^2\cdot \exp\left[-\frac{1}{2}\left(\frac{v^2}{1+\rho} + \frac{w^2}{1-\rho}\right)\right]\frac{dv}{\sqrt{1+\rho}}\frac{dw}{\sqrt{1-\rho}} - \right.$$

$$\left.\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w^2\cdot \exp\left[-\frac{1}{2}\frac{v^2}{1+\rho} + \frac{w^2}{1-\rho}\right)\right]\frac{dv}{\sqrt{1+\rho}}\frac{dw}{\sqrt{1-\rho}}\right\} =$$

$$\frac{\sigma_x\sigma_y}{2}\{(1+\rho) - (1-\rho)\} = \sigma_x\sigma_y\rho, \quad (6)$$

which proves our statement. (Here we have simply observed that the two integrals in { } of (6) give the dispersions of a two-dimensional random variable which is normally distributed with the parameters $(\mu_v, \mu_w, \sigma_v, \sigma_w, \rho_{vw}) = (0, 0, \sqrt{1+\rho}, \sqrt{1-\rho}, 0)$.)

***Exercise 1.** (Cf. Appendix 2). We note that the expression in the exponential of (1) is a homogeneous, quadratic form in $(t_1 - \mu_1) = (t - \mu_x)$ and $(t_2 - \mu_2) = (u - \mu_y)$:

$$-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij}(t_i - \mu_i)(t_j - \mu_j).$$

Comparing with (1) show that the quadratic, symmetric matrix $\boldsymbol{A} = \{a_{rs}\}$ of this form is given by

$$\boldsymbol{A} = \frac{1}{1-\rho^2}\begin{Bmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\sigma_y} \\ -\frac{\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{Bmatrix}. \quad (7)$$

Find the determinant, $|\boldsymbol{A}| \neq 0$, and show that Φ in (1) may also be written

$$d\Phi = \varphi(t, u)\, dt\, du = \frac{\sqrt{|\boldsymbol{A}|}}{(\sqrt{2\pi})^2}\exp\left[-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij}(t_i - \mu_i)(t_j - \mu_j)\right] dt_i\, dt_j =$$

$$\frac{\sqrt{|\boldsymbol{A}|}}{(\sqrt{2\pi})^2}\exp\left[-\frac{1}{2}(\boldsymbol{T} - \mathfrak{M})^* \cdot \boldsymbol{A} \cdot (\boldsymbol{T} - \mathfrak{M})\right] dt_1\, dt_2, \quad (8)$$

in which $\underset{21}{T}$ and $\underset{21}{\mathfrak{M}}$ are the column matrices with the elements t_1, t_2, and μ_1, μ_2, respectively. Finally show that

$$A^{-1} = M, \tag{9}$$

where M is the moment matrix defined in (6.4.17).

*Example 2. (Cf. Appendix 2). From (8) it is now an obvious generalization to define a ν-dimensional normal random variable $(x_1, \cdots, x_\nu)$ as one having the distribution

$$d\Phi = \varphi(t_1, \cdots, t_\nu)\, dt_1 \cdots dt_\nu =$$

$$\frac{\sqrt{|A|}}{(\sqrt{2\pi})^\nu} \exp\left[-\frac{1}{2}\sum_{i=1}^{\nu}\sum_{j=1}^{\nu} a_{ij}(t_i - \mu_i)(t_j - \mu_j)\right] dt_1 \cdots dt_\nu =$$

$$\frac{\sqrt{|A|}}{(\sqrt{2\pi})^\nu} \exp\left[-\frac{1}{2}(T - \mathfrak{M})^* \cdot A \cdot (T - \mathfrak{M})\right] dt_1 \cdots dt_\nu, \tag{10}$$

in which A is a ν-dimensional quadratic, symmetric matrix such that the quadratic form in the exponent is never negative; for it can be shown that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi(t_1, \cdots, t_\nu)\, dt_1 \cdots dt_\nu = 1 \tag{11}$$

$$\mathfrak{M}\{(x_1, \cdots, x_\nu)\} = (\mu_1, \cdots, \mu_\nu) \tag{12}$$

$$M = \{\mathfrak{M}\{(x_r - \mu_r)(x_s - \mu_s)\}\} = A^{-1}. \tag{13}$$

(We note that (13) shows that, if $x_1, \cdots, x_\nu$ are two-by-two uncorrelated, then the x's are, furthermore, also independent.) To prove this we need use only the well-known fact that a symmetric matrix may be transformed into a diagonal one by an orthogonal transformation; i.e., we may introduce ν new random variables, $y_1, \cdots, y_\nu$, by the orthogonal transformation (cf. Exercise 3, §6.4)

$$\underset{\nu 1}{X} - \underset{\nu 1}{\mathfrak{M}} = \underset{\nu\nu}{F} \cdot \underset{\nu 1}{Y}, \tag{14}$$

in which F is orthogonal,

$$F \cdot F^* = E, \quad \text{i.e.,} \quad |F| = \pm 1, \quad \text{and} \quad F^* = F^{-1}, \tag{15}$$

and is determined such that

$$F^* \cdot A \cdot F = F^{-1} \cdot A \cdot F = D, \quad \text{i.e.,} \quad |D| = |F^{-1}||F||A| = |A|. \tag{16}$$

Here D is a diagonal matrix, the diagonal elements of which we shall

denote $1/d_1^2, \cdots, 1/d_\nu^2$. Since from (14) the Jacobian (4.14.2) $\dfrac{\partial(t_1, \cdots, t_\nu)}{\partial(u_1, \cdots, u_\nu)} = |\boldsymbol{F}| = \pm 1$ and from (16) $\sqrt{|\boldsymbol{A}|} = \sqrt{|\boldsymbol{D}|} = 1/d_1 \cdots 1/d_\nu$, we obtain from (4.14.3) and (7.1.2)

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi(t_1, \cdots, t_\nu)\, dt_1 \cdots dt_\nu =$$

$$\frac{\sqrt{|\boldsymbol{D}|}}{(\sqrt{2\pi})^\nu} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\boldsymbol{U}^* \cdot \boldsymbol{D} \cdot \boldsymbol{U}\right] du_1 \cdots du_\nu =$$

$$\prod_{i=1}^{\nu} \left(\frac{1}{\sqrt{2\pi}\, d_i} \int_{-\infty}^{\infty} \exp\left[-\frac{u_i^2}{2d_i^2}\right] du_i\right) = 1, \quad (17)$$

which proves (11). To prove (12) we note that (17) shows that

$$\mathfrak{M}\{y_r\} = \frac{1}{(\sqrt{2\pi})^\nu\, d_1 \cdots d_\nu} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_r \cdot$$

$$\exp\left[-\frac{1}{2}\sum_{i=1}^{\nu} \frac{u_i^2}{d_i^2}\right] du_1 \cdots du_\nu = 0, \quad r = 1, 2, \cdots, \nu; \quad (18)$$

i.e., $\mathfrak{M}\{\boldsymbol{Y}\} = 0$. From (18) and (14) the result (12) follows at once. Furthermore, (17) shows that $\boldsymbol{M}^{(Y)}$ is a diagonal matrix with the diagonal elements $d_1^2, \cdots, d_\nu^2$, i.e., from (15) and (16)

$$\boldsymbol{M}^{(Y)} = \boldsymbol{D}^{-1} = \boldsymbol{F}^{-1} \cdot \boldsymbol{A}^{-1} \cdot \boldsymbol{F}. \quad (19)$$

From (19), (14), (15), and (6.4.23) with $\boldsymbol{Y}$ and $\boldsymbol{X}$ interchanged the result (13) follows at once, because

$$\boldsymbol{M}^{(X)} = \boldsymbol{F} \cdot \boldsymbol{M}^{(Y)} \cdot \boldsymbol{F}^* = \boldsymbol{F} \cdot \boldsymbol{F}^{-1} \cdot \boldsymbol{A}^{-1} \cdot \boldsymbol{F} \cdot \boldsymbol{F}^{-1} = \boldsymbol{A}^{-1}. \quad (20)$$

*Exercise 2. (Cf. Appendix 2). Show that (10) has the characteristic function

$$\chi_{x_1, \cdots, x_\nu}(t_1, \cdots, t_\nu) = \exp\left[i\sum_{k=1}^{\nu} \mu_k t_k - \frac{1}{2}\sum_{k=1}^{\nu}\sum_{l=1}^{\nu} (\boldsymbol{A}^{-1})_{kl} t_k t_l\right] =$$

$$\exp\left[i\mathfrak{M}^* \cdot \boldsymbol{T} - \frac{1}{2}\boldsymbol{T}^* \cdot \boldsymbol{A}^{-1} \cdot \boldsymbol{T}\right] \quad (21)$$

(cf. Example 3, §5.5). If we know (21), show that equations (11)–(13) follow (cf. (5.2.27)). From (21) it follows, furthermore, that the marginal distribution of each x_k, $k = 1, 2, \cdots, \nu$, is normal with the parameters μ_k, σ_k (cf. Exercise 1, §5.5). Thus the characteristic function saves us from directly evaluating the complicated integrals necessary for this result.

For a more detailed discussion of the many-dimensional normal distribution we must refer to textbooks in statistics, e.g., that of Cramér.

*§ 7.7. We shall in this and the following four topics deduce from the normal distribution some other distributions which are very important in the practical applications of the normal distribution (cf. § 11.8–§ 11.17). For later applications it will be convenient to generalize the normal distribution somewhat (cf. Example 2, § 10.3). Let x be a random variable and $y = G(x)$ a function for which $-\infty < y < \infty$, $a \leqq x \leqq b$ $(-\infty \leqq a,\ b \leqq \infty)$ and $G'(x) \gtrless 0$ for all x in the region of definition. We now assume that y is normally distributed, which means (cf. § 4.8) that x has the distribution[1]

$$d\Phi_x = \varphi(x)\,dx = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(G(x) - \mu)^2\right] \left|\frac{dG(x)}{dx}\right| dx. \quad (1)$$

If $G(x) = x$, (1) simply reduces to the usual normal distribution.

Exercise 1. Show that

$$\mathfrak{M}\{G(x)\} = \mu, \qquad \sigma^2\{G(x)\} = \sigma^2. \quad (2)$$

Next we consider n independent variables $x_1, \cdots, x_n$, all having the same distribution (1) (in the later applications $x_1, \cdots, x_n$ will denote n independent observations of one and the same variable x). The joint distribution of $x_1, \cdots, x_n$ is then given by

$$d\Phi = \varphi(x_1, \cdots, x_n)\,dx_1 \cdots dx_n =$$

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(G(x_i) - \mu)^2\right] \prod_{i=1}^{n} |G'(x_i)|\,dx_1 \cdots dx_n. \quad (3)$$

Instead of $x_1, \cdots, x_n$ we introduce n new random variables defined by

$$m = \overline{G(x)} = \frac{G(x_1) + \cdots + G(x_n)}{n} \qquad (-\infty < m < \infty) \quad (4)$$

$$q = \sqrt{\sum_{i=1}^{n} v_i^2} \qquad (0 \leqq q < \infty) \quad (5)$$

$$v_i = G(x_i) - m$$

$$u_i = \frac{G(x_i) - m}{q} \qquad (-1 < u_i < 1) \quad i = 1, 2, \cdots, n. \quad (6)$$

[1] For convenience we use in § 7.7–7.11 the same letter for a random variable and for the variable in the distribution function.

Exercise 2. Show that

$$\mathfrak{M}\{m\} = \mu, \qquad \sigma\{m\} = \frac{\sigma}{\sqrt{n}}. \tag{7}$$

Next show that

$$\sum_{i=1}^{n} u_i = 0, \qquad \sum_{i=1}^{n} u_i^2 = 1 \tag{8}$$

and that this implies that $|u_i| < 1$, i.e., that $|u_i| = 1$ is excluded. Because of the two algebraic relations (8) only $n - 2$ of the u's are free variables. Let us consider $u_1, u_2, \cdots, u_{n-2}$ as the free variables. Thus together with m and q we have introduced in all n new variables. Show that for the Jacobian determinant (4.14.2)

$$\frac{\partial(x_1, \cdots, x_n)}{\partial(m, q, u_1, \cdots, u_{n-2})} = q^{n-2} f(u_1, u_2, \cdots, u_{n-2}; n), \tag{9}$$

in which f is a certain function of the u's alone, which we need not work out. Finally show that

$$\sum_{i=1}^{n} (G(x_i) - \mu)^2 = n(m - \mu)^2 + q^2. \tag{10}$$

As a result of Exercise 2 we find from (4.14.3) that expressed in the new variables our distribution is

$$d\Phi = \left\{\frac{\sqrt{n}}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{n(m-\mu)^2}{2\sigma^2}\right] dm\right\} \cdot \left\{\frac{1}{\left(\frac{n-3}{2}\right)!\, 2^{(n-3)/2}} \left(\frac{q}{\sigma}\right)^{n-2} \exp\left[-\frac{q^2}{2\sigma^2}\right] \frac{dq}{\sigma}\right\} \cdot \tag{11}$$

$$\{\text{const } f(u_1, \cdots, u_{n-2}; n)\, du_1 \cdots du_{n-2}\},$$

where the constants have been chosen so as to normalize each { } separately (check this by means of Appendix 1). Equation (11) shows that $\boldsymbol{m}$ and $\boldsymbol{q}$ are independent variables, that $\boldsymbol{m}$ is normally distributed with the parameters μ and $\sigma/\sqrt{n}$, and that $\boldsymbol{q}$ has the distribution, first found by Helmert,

$$d\Phi = \varphi(q)\, dq = \frac{1}{\left(\frac{f-2}{2}\right)!\, 2^{(f-2)/2}} \left(\frac{q}{\sigma}\right)^{f-1} \exp\left[-\frac{q^2}{2\sigma^2}\right] \frac{dq}{\sigma}$$

$$(0 \leqq q < \infty), \quad f = n - 1. \tag{12}$$

Here we have introduced the so-called **degree of freedom** $f = n - 1$ instead of n, which refers to the fact that q^2 is a sum of squares of the

n variables $v_1, \cdots, v_n$, subject to one constraint given in (8) so that only $n - 1$ of the v's are free variables.

Exercise 3. Show by means of Appendix 1 that

$$\mathfrak{M}\{q^2\} = f\sigma^2, \quad \mathfrak{M}\{q^4\} = (f+2)f\sigma^4, \quad \text{i.e.,} \quad \sigma^2\{q^2\} = 2f\sigma^4. \tag{13}$$

Next show that

$$\mathfrak{M}\{q\} = \frac{\left(\frac{f-1}{2}\right)!}{\left(\frac{f-2}{2}\right)!}\sqrt{2}\,\sigma \sim \sqrt{f-\frac{1}{2}}\,\sigma, \tag{14}$$

i.e.,

$$\sigma^2\{q\} = \left(f - 2\left(\frac{\left(\frac{f-1}{2}\right)!}{\left(\frac{f-2}{2}\right)!}\right)^2\right)\sigma^2 \sim \frac{\sigma^2}{2} \tag{15}$$

(use Stirling's formula for the approximate formulae).

Exercise 4. Show that the q-distribution (12) contains some previous distributions: (1) for $f = 1$ the distribution of $|x|$ when x is normally distributed with the parameters 0, σ; (2) for $f = 2$ the distribution (7.4.5); (3) for $f = 3$ the Maxwell-Boltzmann distribution (4.8.2) (cf. Exercise 3, §7.4).

Exercise 5. Instead of q it is often convenient to introduce

$$s = \frac{q}{\sqrt{f}}, \qquad f = n - 1. \tag{16}$$

Show that

$$\mathfrak{M}\{s^2\} = \sigma^2, \qquad \sigma^2\{s^2\} = \frac{2\sigma^4}{f} \tag{17}$$

and

$$\mathfrak{M}\{s\} = \frac{\left(\frac{f-1}{2}\right)!}{\left(\frac{f-2}{2}\right)!}\sqrt{\frac{2}{f}}\,\sigma \sim \sqrt{1-\frac{1}{2f}}\,\sigma \tag{18}$$

$$\sigma^2\{s\} = \left(1 - \frac{2}{f}\left(\frac{\left(\frac{f-1}{2}\right)!}{\left(\frac{f-2}{2}\right)!}\right)^2\right)\sigma^2 \sim \frac{\sigma^2}{2f}. \tag{19}$$

*§7.8. In many applications it is more convenient to work with the distribution of q^2 rather than that of q. Furthermore, it is convenient to consider the variable q^2/σ^2, which is denoted by[1]

[1] Although the application of a Greek letter for a quantity which is not a parameter, but a variable, is against our fundamental principle for notations, and although

$$\chi^2 = \frac{q^2}{\sigma^2}. \tag{1}$$

Exercise 1. Show that χ^2 has the distribution

$$d\Phi = \varphi(\chi^2)\, d(\chi^2) = \frac{1}{\left(\frac{f-2}{2}\right)!\, 2^{f/2}} (\chi^2)^{(f-2)/2} \exp\left[-\frac{1}{2}\chi^2\right] d(\chi^2)$$

with

$$(0 \leqq \chi^2 < \infty) \tag{2}$$

$$\mathfrak{M}\{\chi^2\} = f, \quad \mathfrak{M}\{\chi^4\} = (f+2)f, \quad \text{i.e.,} \quad \sigma^2\{\chi^2\} = 2f. \tag{3}$$

Putting $P(\chi^2) = \int_{\chi^2}^{\infty} \varphi(\chi^2)\, d(\chi^2)$ we give in Table V χ^2 as a function of f for $P = 95\%, 10\%, 5\%, 1\%$, and 0.1%.[1] It may be shown that for large values of f the variable $\sqrt{2\chi^2}$ is approximately normally distributed with the mean $\sqrt{2f-1}$ and the dispersion 1. The χ^2-distribution has the important property that, if ${\chi_1}^2$ and ${\chi_2}^2$ are independent and both have the distribution (2) with the degrees of freedom f_1 and f_2, respectively, then $\chi^2 = {\chi_1}^2 + {\chi_2}^2$ has the distribution (2) with $f = f_1 + f_2$.

Exercise 2. Verify this. (Introduce in the joint distribution of ${\chi_1}^2$ and ${\chi_2}^2$ the new variables χ^2 and p given by ${\chi_1}^2 = p\chi^2$, ${\chi_2}^2 = (1-p)\chi^2$, and integrate over p from 0 to 1 (cf. (4.14.6).)

Exercise 3. If $y_1, \cdots, y_n$ are independent and each normally distributed with $\mu = 0$ and $\sigma = 1$, show that

$$\chi^2 = {y_1}^2 + {y_2}^2 + \cdots + {y_n}^2 \tag{4}$$

has the distribution (2) with the degree of freedom $f = n$. Then show that under the assumptions of §7.7

$${\chi_0}^2 = \frac{{q_0}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} (G(x_i) - \mu)^2 \tag{5}$$

has the same distribution as

$${\chi_1}^2 = \frac{q^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} (G(x_i) - m)^2, \tag{6}$$

the only difference being that ${\chi_0}^2$ has $f = n$, but ${\chi_1}^2$ has $f = n-1$.

*§ 7.9. Let y and q be independent variables, and let y be a normally distributed variable with $\mathfrak{M}\{y\} = 0$ and $\sigma\{y\} = a\sigma$ (a, a constant); and let q be a variable with the distribution (7.7.12) with a certain degree of freedom f. Then the normalized variable corresponding to

---

the notation χ^2 instead of χ is rather unhandy, we have not had the courage to apply another symbol, since χ^2 seems to be universally standardized.

[1] For a more detailed table see Fisher and Yates, *Statistical Tables*, Table IV.

y is $y/a\sigma$ (cf. Exercise 2, §5.3). We now replace σ by s given in (7.7.16) and call the result t:

$$t = \frac{y}{as} = \frac{\sqrt{f}}{a}\frac{y}{q}. \tag{1}$$

To obtain the distribution of t we write down the joint distribution of y and q. Next we introduce as new variables t and q instead of y and q and obtain from (4.14.3)

$$d\Phi = \varphi(t, q)\, dt\, dq = \frac{1}{\sqrt{2\pi}\,\sqrt{f}\left(\frac{f-2}{2}\right)!\,2^{\frac{f-2}{2}}}\left(\frac{q}{\sigma}\right)^f \exp\left[-\left(1+\frac{t^2}{f}\right)\frac{q^2}{2\sigma^2}\right] dt\,\frac{dq}{\sigma} \tag{2}$$

(check). Integrating (2) over all values of q we get the marginal distribution of t, by introducing $\left(1+\frac{t^2}{f}\right)^{\frac{1}{2}}\frac{q}{\sqrt{2}\,\sigma}$ as a new integration variable,

$$d\Phi = \varphi(t)\, dt = \frac{1}{\sqrt{\pi}\,\sqrt{f}}\frac{\left(\frac{f-1}{2}\right)!}{\left(\frac{f-2}{2}\right)!}\frac{dt}{\left(1+\frac{t^2}{f}\right)^{\frac{f+1}{2}}} \qquad (-\infty < t < \infty) \tag{3}$$

(check this result by means of Appendix 1). This distribution is called **Student's distribution.** We stress that it is independent of both the parameters μ and σ, depending only on the degree of freedom f. Since $\left(1+\frac{t^2}{f}\right)^{-(f+1)/2} \xrightarrow[f\to\infty]{} \exp\left[-\frac{t^2}{2}\right]$ we shall expect that the t-distribution is very close to the normalized normal distribution (7.1.6) for large values of f.

Exercise 1. Verify this by means of Stirling's formula.

The probability $P(t)$ for $|t| \geqq t$ is given by

$$P(t) = 2\int_t^\infty \varphi_t(t)\, dt. \tag{4}$$

In Table III we have given t as a function of f for $P = 10\%$, 5%, 1%, and 0.1%.[1] It will be seen that for $f \to \infty$ we obtain the corresponding tolerance limits of the normal distribution (cf. Table II).

[1] A more detailed table of the t-distribution is given in Fisher and Yates, *Statistical Tables*, Table III.

Exercise 2. Let $\boldsymbol{m}$ be defined in (7.7.4), $\boldsymbol{q}$ in (7.7.5), and $\boldsymbol{s}$ in (7.7.16). Then show that

$$\boldsymbol{t} = \frac{\boldsymbol{m} - \mu}{\frac{\boldsymbol{s}}{\sqrt{n}}} = \sqrt{n(n-1)}\,\frac{\boldsymbol{m} - \mu}{\boldsymbol{q}} \tag{5}$$

has the t-distribution (3) with $f = n - 1$. Next let x be a variable which is independent of $\boldsymbol{m}$ and of $\boldsymbol{q}$ and which has the distribution (7.7.1). Then show that

$$\boldsymbol{t} = \frac{\boldsymbol{m} - G(x)}{\sqrt{1 + \frac{1}{n}}\,\boldsymbol{s}} = \sqrt{\frac{n(n-1)}{n+1}}\,\frac{\boldsymbol{m} - G(x)}{\boldsymbol{q}} \tag{6}$$

has also the t-distribution (3) with $f = n - 1$.

Exercise 3. Let $x_{11}, x_{12}, \cdots, x_{1n_1}$ and $x_{21}, x_{22}, \cdots, x_{2n_2}$ be two mutually independent set of independent observations of x having the distribution (7.7.1). Let $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ be the quantities corresponding to the two sets as defined in (7.7.4), and let $\boldsymbol{s}_1 = \boldsymbol{q}_1/\sqrt{f_1}$ and $\boldsymbol{s}_2 = \boldsymbol{q}_2/\sqrt{f_2}$ be the corresponding quantities as defined in (7.7.16). Then show that

$$\mathfrak{M}\{\boldsymbol{m}_1 - \boldsymbol{m}_2\} = 0, \qquad \sigma\{\boldsymbol{m}_1 - \boldsymbol{m}_2\} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{7}$$

Next show that the "normalized" variable

$$\boldsymbol{t} = \frac{\boldsymbol{m}_1 - \boldsymbol{m}_2}{\boldsymbol{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \boldsymbol{s} = \frac{\boldsymbol{q}}{\sqrt{f}} = \frac{\sqrt{\boldsymbol{q}_1^2 + \boldsymbol{q}_2^2}}{\sqrt{f_1 + f_2}} \quad \begin{matrix} f_1 = n_1 - 1 \\ f_2 = n_2 - 1, \end{matrix} \tag{8}$$

has the t-distribution (3) with $f = f_1 + f_2$.

*§ 7.10. Let us consider the "relative deviation" of $G(x_i)$, replacing μ by $\boldsymbol{m}$ and again σ by $\boldsymbol{s} = \boldsymbol{q}/\sqrt{n-1}$,

$$r_i = \frac{v_i}{\sqrt{\frac{n-1}{n}}\,\boldsymbol{s}} = \sqrt{n}\,\frac{G(x_i) - \boldsymbol{m}}{\boldsymbol{q}} = \sqrt{n}\,u_i, \qquad i = 1, 2, \cdots, n, \tag{1}$$

where m, v_i, u_i, q, and s are all given in § 7.7. It may be shown directly from (7.7.11) that the marginal distribution of each r_i has the distribution[1]

[1] See Arley, *Danske Vid. Selsk. Mat.-fys. Medd.*, Vol. XVIII, No. 3, 1940. Here also the joint distribution of $r_1, \cdots, r_{n-2}$ and a more detailed table of the r-distribution are given.

$$d\Phi = \varphi(r)\,dr = \frac{1}{\sqrt{\pi}}\,\frac{1}{\sqrt{f+1}}\,\frac{\left(\frac{f-1}{2}\right)!}{\left(\frac{f-2}{2}\right)!}\left(1-\frac{r^2}{f+1}\right)^{(f-2)/2}dr,$$

$$|r| \leqq \sqrt{f+1}, \quad f = n-2. \tag{2}$$

Exercise 1. Show that we have two constraints on the r's

$$\sum_{i=1}^{n} r_i = 0, \qquad \sum_{i=1}^{n} r_i^2 = n, \tag{3}$$

i.e., one more constraint than on the v_i's, which is the reason for putting $f = n-2$ here.

The distribution (2) may also be derived from the t-distribution, as stated in the following exercise.

Exercise 2. Show that t' defined by

$$t' = \frac{\sqrt{n-2}\,r_1}{\sqrt{n-1-r_1^2}} = \frac{v_1}{\sqrt{\frac{1}{n-2}\sum_{i=2}^{n} v_i^2}} \tag{4}$$

has the t-distribution (7.9.3) with $f = n-2$. Next show that, when r_1 increases from $-\sqrt{n-1}$ to $\sqrt{n-1}$, t' will increase from $-\infty$ to ∞ and thus that $P(r_1 \leqq r_1) = P\left(t' \leqq \frac{\sqrt{n-2}\,r_1}{\sqrt{n-1-r_1^2}}\right)$. Finally, inserting the distribution for t' and differentiating, deduce (2).

We note that the r-distribution is, like the t-distribution, independent of both the parameters μ and σ and further that it will be very nearly equal to the normalized normal distribution for large values of f. The probability $P(r)$ of $|r| \geqq r$ is given by

$$P(r) = 2\int_r^{\sqrt{f+1}} \varphi_r(r)\,dr. \tag{5}$$

In Table IV we have given r as a function of f for $P = 10\%$, 5%, 1%, and 0.1%. It will be seen that for $f \to \infty$ we exactly obtain the corresponding tolerance limits of the normal distribution (cf. Table II).

*§ 7.11. Finally we consider two independent variables, q_1 and q_2, both having the q-distribution (7.7.12) with the degrees of freedom f_1 and f_2, respectively. Introducing into their joint distribution the

variable w, called the **variance quotient,**

$$w = \frac{s_1}{s_2} = \sqrt{\frac{f_2}{f_1}}\,\frac{q_1}{q_2} \qquad (0 \leqq w < \infty) \tag{1}$$

instead of q_1, we get from (4.14.3)

$$d\Phi = \varphi(w, q_2)\, dw\, dq_2 =$$

$$\frac{1}{\left(\frac{f_1-2}{2}\right)!\left(\frac{f_2-2}{2}\right)!\,2^{\frac{f-4}{2}}}\left(\frac{f_1}{f_2}\right)^{f_1/2} w^{f_1-1}\left(\frac{q_2}{\sigma}\right)^{f-1}\cdot$$

$$\exp\left[-\left(1+\frac{f_1}{f_2}w^2\right)\frac{{q_2}^2}{2\sigma^2}\right] dw\,\frac{dq_2}{\sigma}, \qquad f = f_1 + f_2. \tag{2}$$

Integrating over q_2 from 0 to ∞ we obtain the marginal distribution of w by introducing $\left(1+\frac{f_1}{f_2}w^2\right)^{\frac{1}{2}}\frac{q_2}{\sqrt{2}\,\sigma}$ as a new integration variable

$$d\Phi = \varphi(w)\, dw = \frac{2\left(\frac{f-2}{2}\right)!}{\left(\frac{f_1-2}{2}\right)!\left(\frac{f_2-2}{2}\right)!}\left(\frac{f_1}{f_2}\right)^{f_1/2}\frac{w^{f_1-1}}{\left(1+\frac{f_1}{f_2}w^2\right)^{f/2}}\,dw \tag{3}$$

(check this result by means of Appendix 1).

Instead of w it is often more convenient to introduce w^2 denoted F.

Exercise 1. Show that the distribution of $F = w^2$ is given by

$$d\Phi = \varphi(w^2)\, d(w^2) = \frac{\left(\frac{f-2}{2}\right)!}{\left(\frac{f_1-2}{2}\right)!\left(\frac{f_2-2}{2}\right)!}\left(\frac{f_1}{f_2}\right)^{f_1/2}\frac{(w^2)^{(f_1-2)/2}}{\left(1+\frac{f_1}{f_2}w^2\right)^{f/2}}\,d(w^2),$$

$$f = f_1 + f_2 \quad (0 \leqq w^2 < \infty). \tag{4}$$

We stress that, just as the t- and the r-distributions are independent of the primary parameters μ and σ, so is the w^2-distribution. The probability $P(w^2)$ of $w^2 \geqq w^2$ is given by

$$P(w^2) = \int_{w^2}^{\infty} \varphi_{w^2}(w^2)\, d(w^2). \tag{5}$$

In Table VI we have given w^2 as a function of f_1 and f_2 for $P = 5\%$.[1]

[1] A more detailed table is given in Fisher and Yates, *Statistical Tables*, Table V (in which $e^{2z} = w^2 = F$).

It is assumed here that the numbers 1 and 2 have been chosen such that $s_1 \geqq s_2$, i.e., $w^2 \geqq 1$. We note that the w^2-distribution contains the χ^2-, and the t-, and thus also the r-distribution as special cases. For $w = \chi/\sqrt{f_1}$ and $f_2 \to \infty$, (4) reduces to (7.8.2) with $f = f_1$. For $w = t$ and $f_1 = 1$, (4) reduces to (7.9.3) with $f = f_2$.

Exercise 2. Check these facts by means of the tables.

8.

LIMIT THEOREMS

§8.1. Let $u = f(t)$ be a non-negative function for which

$$u = f(t) \geqq a > 0 \tag{1}$$

for some constant, a, and for all values of t within a certain region ω. For a random variable $\boldsymbol{x}$ with continuous distribution we then have from (5.2.1)

$$\mathfrak{M}\{f(\boldsymbol{x})\} = \int_{-\infty}^{\infty} f(t)\varphi(t)\,dt \geqq a \int_{\omega} \varphi(t)\,dt = aP(\boldsymbol{x} \text{ in } \omega), \tag{2}$$

i.e.,

$$P(\boldsymbol{x} \text{ in } \omega) \leqq \frac{\mathfrak{M}\{f(\boldsymbol{x})\}}{a}. \tag{3}$$

Exercise 1. Show that (3) also holds for discontinuous distributions.

If in (3) we put $f(\boldsymbol{x}) = (\boldsymbol{x} - \mathfrak{M}\{\boldsymbol{x}\})^2$ and $\sqrt{a} = k$, we obtain from (5.3.1)

$$P(|\boldsymbol{x} - \mathfrak{M}\{\boldsymbol{x}\}| \geqq k) \leqq \frac{\sigma^2\{\boldsymbol{x}\}}{k^2}, \tag{4}$$

which inequality is called **Tschebyscheff's inequality.**

Exercise 2. Find the corresponding inequality for $f(\boldsymbol{x}) = |\boldsymbol{x} - \mathfrak{M}\{\boldsymbol{x}\}|^n$ and $\sqrt[n]{a} = k$.

If we have a sequence of random variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ for which the means $\mu_1, \mu_2, \cdots$ and the dispersions $\sigma_1, \sigma_2, \cdots$ all exist, we find from (4) that, if

$$\mathfrak{M}\{(\boldsymbol{x}_\nu - \mu_\nu)^2\} = \sigma_\nu^2 \xrightarrow[\nu \to \infty]{} 0, \tag{5}$$

then for an arbitrary, but fixed, $\epsilon > 0$

$$P(|\boldsymbol{x}_\nu - \mu_\nu| \geqq \epsilon) \leqq \frac{\sigma_\nu^2}{\epsilon^2} \xrightarrow[\nu \to \infty]{} 0. \tag{6}$$

We then say that the random variable $\boldsymbol{x}_\nu - \mu_\nu$ **converges in proba-**

bility to 0.[1] More generally, if there exists such a random variable x that

$$P(|x_\nu - x| \geqq \epsilon) \underset{\nu \to \infty}{\longrightarrow} 0, \tag{7}$$

we say that x_ν converges in probability to x. This is a very important concept in modern probability. We stress that it must not be confused with ordinary convergence, $x_\nu \underset{\nu \to \infty}{\longrightarrow} x$, where x_ν and x represent observed values of the random variables x_ν and x, respectively. For it is obvious that we can never prove mathematically anything about the behavior of empirical values. What can be proved are only statements regarding the theoretical concepts of our model, never of reality (cf. § 1.1 and § 9.2). A special symbol for "convergence in probability" has, therefore, been introduced:

$$x_\nu \underset{\nu \to \infty}{\overset{P}{\longrightarrow}} x \tag{8}$$

or, more commonly used,

$$x_\nu \underset{\substack{\text{in } p \\ \nu \to \infty}}{\longrightarrow} x. \tag{9}$$

*Exercise 3. Prove that a necessary and sufficient condition for convergence in probability is that, $\epsilon(t)$ being the causal distribution (4.3.3),

$$\Phi_{x_\nu - x}(t) \underset{\nu \to \infty}{\longrightarrow} \epsilon(t) \tag{10}$$

for all fixed $t \neq 0$, or that

$$\Phi_{x_\nu}(t) \underset{\nu \to \infty}{\longrightarrow} \Phi_x(t) \tag{11}$$

for any fixed t-value which is a continuity point of $\Phi(t)$.

*Exercise 4. Let a random variable x with mean μ and dispersion σ exist such that

$$\mathfrak{M}\{(x_\nu - x)^2\} = \sigma^2\{x_\nu - x\} \underset{\nu \to \infty}{\longrightarrow} 0. \tag{12}$$

(We then say that x_ν **converges in mean** to x.) Show that we then also have $x_\nu \underset{\substack{\text{in } p \\ \nu \to \infty}}{\longrightarrow} x$.

Example 1. It is easily seen that we may have $x_\nu \underset{\text{in } p}{\longrightarrow} x$, without one of $\mu_1, \mu_2, \cdots$ and $\sigma_1, \sigma_2, \cdots$ existing. For example, let us consider a sequence of random variables $x_1, x_2, \cdots$ all having the Cauchy distribution (4.4.9) with the parameters $\mu_1 = \mu_2 = \cdots = 0$ and $\alpha_\nu = 1/\nu$, i.e., $\alpha_\nu \underset{\nu \to \infty}{\longrightarrow} 0$. In this case the defining integrals for the means are undetermined and $\sigma_1 = \sigma_2 = \cdots = \infty$; but, as discussed in Example 4, § 4.4, we nevertheless have $\Phi_\nu(t) \underset{\nu \to \infty}{\longrightarrow} \epsilon(t)$ for all

[1] Another expression is **converges stochastically**.

$t \neq 0$, which from Exercise 3 means that $x_\nu \xrightarrow[\nu \to \infty]{\text{in } p} x$, where x has the causal distribution $\epsilon(t)$ given in (4.3.3), i.e., x is a constant $= 0$. (Furthermore we see that $\sigma^2\{x_\nu - x\} = \mathfrak{M}\{(x_\nu - x)^2\} = \mathfrak{M}\{x_\nu^2\} = \infty$, i.e., x_ν does *not* converge in mean to the limit x. Thus convergence in probability is the more general concept and, therefore, of wider applicability.)

In particular, let us in (6) put $x_\nu = r/\nu = f$, where r is the absolute and f the relative frequency among ν independent observations of a certain event which has the probability θ of occurring in a single observation. As shown in §3.7, r is binomially distributed, and from (6.4.14) and (6.4.15) we see that (6) gives for an arbitrary, but fixed, $\epsilon > 0$

$$P(|f - \theta| \geqq \epsilon) \xrightarrow[\nu \to \infty]{} 0, \quad \text{i.e.,} \quad f \xrightarrow[\nu \to \infty]{\text{in } p} \theta, \tag{13}$$

which is called **Bernoulli's theorem:**

The relative frequency f among ν independent observations of an event converges in probability to its probability θ for $\nu \to \infty$, i.e., the probability of f deviating more than ϵ from θ becomes arbitrarily small when ν increases indefinitely.

Next, by putting in (6) for x_ν: $\bar{x} = \frac{1}{\nu}(x_1 + \cdots + x_\nu)$, where $x_1, \cdots, x_\nu$ are ν independent observations of a random variable x with finite mean μ and dispersion σ, (6.4.8) and (6.4.9) show that from (6) we have for an arbitrary, but fixed, $\epsilon > 0$

$$P(|\bar{x} - \mu| \geqq \epsilon) \xrightarrow[\nu \to \infty]{} 0, \quad \text{i.e.,} \quad \bar{x} \xrightarrow[\nu \to \infty]{\text{in } p} \mu, \tag{14}$$

which is called the **law of large numbers.**[1]

The average, $\bar{x}$, of ν independent observations of a random variable x with finite μ and σ converges in probability to μ for $\nu \to \infty$; i.e., the probability of $\bar{x}$ deviating more than ϵ from μ becomes arbitrarily small when ν increases indefinitely.

Example 2. We stress that in order to compare Bernoulli's theorem, or the law of large numbers, with experience, it is f, or $\bar{x}$, which is the random variable; i.e., we have to think of the whole series of ν observations as being only *one* observation of f, or of $\bar{x}$, although a ν-dimensional one. As discussed in Chapter 1 we then have to consider a large number, n, of such "single" observations of f, or $\bar{x}$, i.e.,

[1] See footnote 1, p. 6.

each of the n observations consists itself of ν observations which give, however, only one numerical value f of $\boldsymbol{f}$, or $\bar{x}$ of $\bar{\boldsymbol{x}}$. The relative number n'/n of these n f-values, or $\bar{x}$-values, for which $|f - \theta| \geqq \epsilon$, or $|\bar{x} - \mu| \geqq \epsilon$, will then be expected to lie nearer to the theoretical value, the probability, the larger the value of n for *fixed* value of ν. (That is, if ν itself is very large, the theoretical value is practically 0 as given in (13), or (14).) In order not to confuse the meanings of n and ν we have denoted these two quantities by different letters.

It is important to know that for the law of large numbers it is essential that μ exists in the strict sense, i.e., that the defining sum or integral is absolutely convergent, as shown by the following example.

Example 3. Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_\nu$ have Cauchy's distribution with the same parameters μ and α (cf. (4.4.9)), in which case the integral for the mean is *not* absolutely convergent (cf. Example 9, §5.1). As shown in problem 41, or in Example 2, §7.5, $\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_\nu$ then also has Cauchy's distribution, but with the parameters $\nu\mu$ and $\nu\alpha$. For $\bar{\boldsymbol{x}} = 1/\nu \cdot (\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_\nu)$ it then readily follows that $\bar{\boldsymbol{x}}$ has Cauchy's distribution with the same parameters as each of the $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_\nu$ themselves, viz., μ and α (check). Consequently

$$P(|\bar{\boldsymbol{x}} - \mu| \geqq \epsilon) = \frac{2}{\pi\alpha} \int_{\mu+\epsilon}^{\infty} \frac{dt}{1 + \dfrac{(t-\mu)^2}{\alpha^2}} \tag{15}$$

is independent of ν, and thus $\bar{\boldsymbol{x}}$ does *not* converge in probability to μ.

It might also be thought that $\sigma < \infty$ is a necessary condition for the law of large numbers. However, this is not the case since *the law of large numbers may be proved solely under the condition that μ exists in the strict sense* (Khintchine's theorem).

***Example 4.** This may easily be shown by using characteristic functions. From the definition of $\chi_x(t)$ it follows that, if μ exists, then, for $t \to 0$, $\chi_x(t) = 1 + \mu i t + o(t)$, where $\dfrac{o(t)}{t} \xrightarrow[t\to 0]{} 0$. From (6.2.8) we then have for any fixed t

$$\chi_{\bar{x}}(t) = \left(\chi_x\left(\frac{t}{\nu}\right)\right)^\nu = \left(1 + \mu i \frac{t}{\nu} + o\left(\frac{1}{\nu}\right)\right)^\nu \xrightarrow[\nu\to\infty]{} e^{i\mu t}, \tag{16}$$

which is just the characteristic function of $\epsilon(t - \mu)$ (cf. (5.2.16)). From the convergence theorem stated in §5.2, p. 62, together with Exercise 3 it follows that $\bar{\boldsymbol{x}} \xrightarrow[\nu\to\infty]{\text{in } p} \mu$, Q.E.D.

§8.2. The normal distribution has the important property of being the limit of many other distributions for certain limit processes (cf. §§7.8, 7.9, 7.10, and 10.10). Here we shall only show this in one special, but important case, viz., that of the binomial distribution (cf. Example 2, §4.3). For large values of ν it may be shown that

$$\varphi_i = \binom{\nu}{i} \theta^i (1-\theta)^{\nu-i} = \frac{1}{\sqrt{2\pi\nu\theta(1-\theta)}} \exp\left[-\frac{(i-\nu\theta)^2}{2\nu\theta(1-\theta)}\right] + \frac{R_\nu}{\nu}, \tag{1}$$

where R_ν is numerically smaller than a certain number which is independent of ν. This is sometimes called **Laplace's formula**, although it was actually deduced by Bernoulli. It holds under the assumption that $\sigma\{x\} = \sqrt{\nu\theta(1-\theta)} \gg 1$ and for such values of i that the normalized variable

$$l = \frac{i - \nu\theta}{\sqrt{\nu\theta(1-\theta)}} \tag{2}$$

is bounded. From (1) it follows that the probability of $l_1 \leqq l \leqq l_2$, where l_1 and l_2 are independent of ν, is

$$P(l_1 \leqq l \leqq l_2) = \frac{1}{\sqrt{2\pi\nu\theta(1-\theta)}} \sum \exp\left[-\frac{l^2}{2}\right] + \sum \frac{R_\nu}{\nu}, \tag{3}$$

in which the summations have to be extended over those values of i for which $l_1 \leqq l \leqq l_2$. The number of such i-values is in any case smaller than $(l_2 - l_1)\sqrt{\nu\theta(1-\theta)} + 1$ (the second term is due to the possibility that both end points l_1 and l_2 may correspond to integer values of i). Thus $\Sigma R_\nu/\nu \longrightarrow_{\nu\to\infty} 0$. In the first term on the right-hand side of (3) we have a sum of values of the function $\exp[-l^2/2]$ for which from (2) the distance between two consecutive l-values is $1/\sqrt{\nu\theta(1-\theta)}$. From the definition of the definite integral we thus have

$$P(l_1 \leqq l \leqq l_2) \longrightarrow_{\nu\to\infty} \frac{1}{\sqrt{2\pi}} \int_{l_1}^{l_2} e^{-t^2/2}\, dt, \tag{4}$$

which is called **de Moivre's theorem.** Therefore, for large values of ν, l and, consequently, x and $f = x/\nu$ are approximately normally distributed.

Exercise 1. Show that Bernoulli's theorem (8.1.13) is contained in this result.

***Example.** Laplace's formula may be proved by means of Stirling's formula (cf. Appendix 1). Inserting this formula for $\nu!$, $i!$,

and $(\nu - i)!$ on the left-hand side of (1) we find, after some calculation,

$$\binom{\nu}{i}\theta^i(1-\theta)^{\nu-i} =$$

$$\frac{1}{\sqrt{2\pi\nu\theta(1-\theta)}}\left(\frac{\nu\theta}{i}\right)^{i+\frac{1}{2}}\left(\frac{\nu(1-\theta)}{\nu-i}\right)^{\nu-i+\frac{1}{2}}\exp\left[\frac{1}{12}\left(\frac{\theta_1}{\nu}-\frac{\theta_2}{i}-\frac{\theta_3}{\nu-i}\right)\right] =$$

$$\frac{1}{\sqrt{2\pi\nu\theta(1-\theta)}}e^{-z} \quad (5)$$

where introducing l from (2), z is given by

$$z = \left(i+\frac{1}{2}\right)\ln\left(\frac{i}{\nu\theta}\right) + \left(\nu-i+\frac{1}{2}\right)\ln\left(\frac{\nu-i}{\nu(1-\theta)}\right) -$$

$$\frac{1}{12}\left(\frac{\theta_1}{\nu}-\frac{\theta_2}{i}-\frac{\theta_3}{\nu-i}\right) =$$

$$\left(\nu\theta + l\sqrt{\nu\theta(1-\theta)} + \frac{1}{2}\right)\ln\left(1+\frac{l(1-\theta)}{\sqrt{\nu\theta(1-\theta)}}\right) +$$

$$\left(\nu(1-\theta) - l\sqrt{\nu\theta(1-\theta)} + \frac{1}{2}\right)\ln\left(1-\frac{l\theta}{\sqrt{\nu\theta(1-\theta)}}\right) -$$

$$\frac{1}{12}\left(\frac{\theta_1}{\nu}-\frac{\theta_2}{i}-\frac{\theta_3}{\nu-i}\right) \quad (6)$$

(check). For sufficiently large values of ν both $\frac{l(1-\theta)}{\sqrt{\nu\theta(1-\theta)}}$ and $\frac{l\theta}{\sqrt{\nu\theta(1-\theta)}}$ are numerically smaller than $\frac{1}{4}$, l lying between the limits l_1 and l_2, which are independent of ν. In that case we have from Taylor's formula $\ln(1+x) = x - \frac{x^2}{2} + \theta x^3$, $|\theta| < 1$. Introducing this into (6) we find after some calculation

$$z = \frac{l^2}{2} + \frac{r_\nu}{\sqrt{\nu}}, \quad (7)$$

where r_ν is a number which depends on ν, θ, and l but is numerically smaller than a certain number which is independent of ν and l. From (7) we readily obtain (1).

*Exercise 2. Show by means of partial integration that we have the following exact expressions:

$$\sum_{i=r}^{\nu} \binom{\nu}{i} \theta^i (1-\theta)^{\nu-i} = \frac{\nu!}{(r-1)!(\nu-r)!} \int_0^\theta t^{r-1}(1-t)^{\nu-r}\, dt$$

$$(r = 1, 2, \cdots, \nu) \quad (8)$$

$$\sum_{i=0}^{r} \binom{\nu}{i} \theta^i (1-\theta)^{\nu-i} = \frac{\nu!}{r!(\nu-r-1)!} \int_\theta^1 t^r (1-t)^{\nu-r-1}\, dt$$

$$(r = 0, 1, \cdots, \nu - 1). \quad (9)$$

The integrals occurring on the right-hand sides are called incomplete B-functions and have been tabulated.[1]

*§ 8.3. The result (8.2.4) is only a special case of a much more general limit theorem. In § 7.5 we have seen that the sum of ν independent, normally distributed, random variables is also normally distributed. However, if only ν is sufficiently large the sum will in general be approximately normally distributed even though the x's are not. This is called the **central limit theorem:**

Let $x_1, x_2, \cdots$ be a series of independent random variables having arbitrary distributions for which the means, $\mu_1, \mu_2, \cdots$, and the dispersions $\sigma_1, \sigma_2, \cdots$, all exist. We form the new random variables $\sum_{i=1}^{\nu} x_i$ and normalize them by putting

$$y_\nu = \frac{\sum_{i=1}^{\nu} x_i - \sum_{i=1}^{\nu} \mu_i}{\left(\sum_{i=1}^{\nu} \sigma_i^2\right)^{\frac{1}{2}}}, \quad \nu = 1, 2, \cdots \quad (1)$$

$$\mathfrak{M}\{y_\nu\} = 0, \quad \sigma\{y_\nu\} = 1,$$

(check). Under very general conditions we then have that

$$\Phi_{y_\nu}(t) \xrightarrow[\nu \to \infty]{} \Psi(t), \quad (2)$$

in which $\Psi(t)$ is the normalized, normal distribution function. Thus for large values of ν, y_ν and, consequently, also $x_1 + \cdots + x_\nu$ are approximately normally distributed. As a *sufficient* condition for (2) we may mention the following: (2) holds true if there exist two numbers

[1] K. Pearson, *Tables of the Incomplete B-function*, London, 1934.

m and M such that

$$\sigma_i^2 = \mathfrak{M}\{(x_i - \mu_i)^2\} > m > 0 \quad \text{and} \quad \mathfrak{M}\{|x_i - \mu_i|^3\} < M, \quad \text{for all } i = 1, 2, \cdots. \tag{3}$$

We see that (3) implies that

$$\sum_{i=1}^{\nu} \sigma_i^2 \xrightarrow[\nu \to \infty]{} \infty, \tag{4}$$

which may be said to express that no single x_i is dominating.

If (4) is not satisfied, and, therefore, (3) is not satisfied, it is necessary if (2) is to hold that *all* the x_i's be normally distributed (which implies $\Phi_{y_\nu}(t) = \Psi(t)$ for all ν).

Obviously (3) is satisfied if $x_1, x_2, \cdots$ all have one and the same distribution: Thus, since in the binomial distribution we may write $x = \sum_{i=1}^{\nu} x_i$, where x_i all have the same distribution (cf. Example 1, §6.4), de Moivre's theorem (8.2.4) is contained in the central limit theorem. This theorem holds for even weaker conditions than (3), but we shall here give neither the exact conditions nor the proofs.[1] The theorem also holds for certain classes of *dependent* random variables.[2]

Example 1. For the case that $x_1, x_2, \cdots$ all have one and the same distribution the result (2) may easily be proved by means of the characteristic functions. Let the distribution function of the normalized variable $\frac{x_i - \mu_i}{\sigma_i} = z_i$ be $\Phi(t)$, $i = 1, 2, \cdots$ and the corresponding characteristic function be $\chi(t)$. From the definition of $\chi(t)$ and the fact that $\mu = 0$ and $\sigma = 1$ for $\Phi(t)$ we have for $t \to 0$ $\chi(t) = 1 - \frac{1}{2}t^2 + o(t^2)$, where $\frac{o(t^2)}{t^2} \xrightarrow[t \to 0]{} 0$. Thus $\chi\left(\frac{t}{\sqrt{\nu}}\right) = 1 - \frac{t^2}{2\nu} + o\left(\frac{1}{\nu}\right)$. However, from (6.2.8), the characteristic function of $y_\nu = \frac{z_1 + \cdots + z_\nu}{\sqrt{\nu}}$ is then

$$\chi_{y_\nu}(t) = \left(\chi\left(\frac{t}{\sqrt{\nu}}\right)\right)^\nu = \left(1 - \frac{t^2}{2\nu} + o\left(\frac{1}{\nu}\right)\right)^\nu \xrightarrow[\nu \to \infty]{} \exp\left[-\frac{t^2}{2}\right], \tag{5}$$

[1] See, e.g., Cramér, *Mathematical Methods of Statistics;* Cramér, *Random Variables* Chapter VI, or Khintchine, *Asymptotische Gesetze der Wahrscheinlichkeitsrechnung.*

[2] See especially, Lévy, *Théorie de l'addition des variables aléatoires.*

which is just the characteristic function of $\Psi(t)$ (cf. (5.2.21)). From the convergence theorem stated in §5.2, p. 62, (2) follows at once, Q.E.D.

Example 2. It is of interest to observe that one may construct an example showing that, if (4) is *not* fulfilled, (2) need not be true. Let us consider a random variable x which has Laplace's distribution (4.4.10) with the parameters $\mu = 0$ and $\alpha = 1$, i.e., from Exercise 2, §5.1, $\mathfrak{M}\{x\} = \mu = 0$ and from (5.3.15) $\sigma\{x\} = \sqrt{2}\,\alpha = \sqrt{2}$. Next we put $x_1 = \frac{2}{\pi}\frac{1}{1}x$, $x_2 = \frac{2}{\pi}\frac{1}{3}x$, $x_3 = \frac{2}{\pi}\frac{1}{5}x$, $\cdots$. Thus $\sum_{i=1}^{\nu}\mu_i = 0$ and $\sum_{i=1}^{\nu}\sigma_i^2 = \frac{8}{\pi^2}\sum_{i=1}^{\nu}\frac{1}{(2i-1)^2} \xrightarrow[\nu\to\infty]{} \frac{8}{\pi^2}\sum_{i=1}^{\infty}\frac{1}{(2i-1)^2} = \frac{8}{\pi^2}\zeta(2)(1-2^{-2}) = 1$, where $\zeta(z)$ is the zeta function of Riemann.[1] Thus (4) is not satisfied. To work out $\lim_{\nu\to\infty}\Phi_{y_\nu}(t) = \Phi_y(t)$, $y = \lim_{\nu\to\infty} y_\nu$, $y_\nu = \sum_{i=1}^{\nu} x_i$, $\mathfrak{M}\{y\} = 0$ and $\sigma\{y\} = 1$, we use the properties of the characteristic functions. From (5.2.23) $\chi_x(t) = 1/(1+t^2)$. Therefore, from Exercise 2, §6.2, and the convergence theorem of Example 2, §5.2, we have

$$\chi_y(t) = \lim_{\nu\to\infty}\chi_{x_1}(t)\cdots\chi_{x_\nu}(t) = \prod_{i=1}^{\infty}\frac{1}{1+\left(\frac{2}{\pi}\frac{t}{2i-1}\right)^2} = \frac{1}{\cosh t} = \frac{2}{e^t+e^{-t}}.$$

From (5.2.14) we then obtain

$$\varphi_y(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iut}\frac{2}{e^t+e^{-t}}\,dt = \frac{1}{\exp[\frac{1}{2}\pi u] + \exp[-\frac{1}{2}\pi u]} = \frac{1}{2\cosh(\frac{1}{2}\pi u)}, \quad (6)$$

where the integral may be worked out by means of complex integration around a rectangle of height π and indefinitely great breadth, one of whose sides is the real t-axis. Obviously (6) differs from the normal distribution.

§8.4. In §8.2 we considered the limit of the binomial distribution for $\nu \to \infty$ and θ held constant. We shall now consider the limit if

[1] For this and the following calculations, see, e.g., Whittaker and Watson, *Modern Analysis*, Cambridge, 1935.

instead of θ we keep $\mathfrak{M}\{x\} = \mu = \nu\theta$ constant for $\nu \to \infty$, i.e., $\theta \underset{\nu\to\infty}{\longrightarrow} 0$. We then obtain from (4.3.4) introducing μ instead of θ in φ_i

$$\lim_{\nu\to\infty} \varphi_i = \lim_{\nu\to\infty} \binom{\nu}{i} \left(\frac{\mu}{\nu}\right)^i \left(1 - \frac{\mu}{\nu}\right)^{\nu-i} =$$

$$\frac{\mu^i}{i!} \lim_{\nu\to\infty} 1 \cdot \left(1 - \frac{1}{\nu}\right) \cdots \left(1 - \frac{i-1}{\nu}\right) \left(1 - \frac{\mu}{\nu}\right)^{-i} \cdot \left(1 - \frac{\mu}{\nu}\right)^{\nu} =$$

$$\frac{\mu^i}{i!} e^{-\mu}, \quad (1)$$

since $\lim_{x\to 0} (1 + x)^{1/x} = e$. Thus we find that for μ kept constant the binomial distribution converges when $\nu \to \infty$ to Poisson's distribution. If both $\nu \gg i$ and $\nu \gg \mu = \nu\theta$, i.e., $\theta \ll 1$, we find from (1) the formula corresponding to (8.2.1)

$$\varphi_i = \binom{\nu}{i} \theta^i (1 - \theta)^{\nu-i} = \frac{\mu^i e^{-\mu}}{i!} + \frac{R_\nu}{\nu}, \qquad \mu = \nu\theta, \quad (2)$$

in which R_ν is again numerically smaller than a certain number which is independent of ν. Equation (2) is called **Poisson's formula** or, owing to the condition $\theta \ll 1$, the **law of small numbers.** We have previously discussed examples of random variables which have Poisson's distribution (cf. Examples 4 and 5, §4.3).

Example. If μ is very large we must expect Poisson's formula to go over into that of Laplace with $\nu\theta = \mu$ and $\theta = 0$, viz.,

$$\varphi_i = \frac{1}{\sqrt{2\pi\mu}} \exp\left[-\frac{(i-\mu)^2}{2\mu}\right] + \frac{R_\nu}{\nu}, \quad (3)$$

which may also be verified directly from (2) by means of Stirling's formula.

9.

THE RELATION OF THE THEORY OF PROBABILITY TO EXPERIENCE AND ITS PRACTICAL IMPORTANCE

§ 9.1. We have now finished our discussion of the mathematical theory, and before we proceed to discuss some of its applications we shall shortly discuss the two questions: (I) How can the theory be tested by experience? (II) Where is probability applied and of what use is it?

1. THE RELATION TO EXPERIENCE

§ 9.2. All the theorems deduced so far have been of a purely mathematical nature and do not, of course, *prove* anything about what really happens in nature, just as, e.g., in mechanics we cannot prove that the planets will actually follow the paths calculated. What can be proved mathematically are only statements regarding our model of nature, never statements regarding nature itself. Thus Bernoulli's theorem does not prove that the empirical relative frequency will be very nearly equal to its theoretical value, the probability, but only that this is highly probable. However, in the course of time we have collected a large body of empirical material, which shows what no mathematical argument can ever prove that, if an event has a negligibly small probability of occurring, we can assume with a certainty sufficiently high for all practical purposes that the event will not occur in a single observation.

Therefore, by any comparison between theory and experience we shall as a main criterion for agreement demand:

A. If an event has a negligibly small probability of occurring, it should not, "practically speaking," occur in a single observation.

When we do not demand that the event shall never occur at all it is because even the most improbable event may occur. Thus, e.g., if we should prepare the annual statement of a large bank we may, of course, try to take a shortcut and guess all the numbers, and, in principle, it is possible that all the numbers will come out correctly—but

no human being will deny that this event is so unlikely to occur that we may safely neglect its occurrence for all practical purposes (cf. problem 4).

Example 1. In applying the criterion **A** we must be somewhat careful in those cases in which we have only a finite number of equally probable possibilities. For example, since all combinations in a bridge play are equally probable, we might expect that such a combination wherein at least one player gets 13 cards of the same color is just as possible as any other combination, in spite of the fact that the probability of the first event is only $1/(4 \times 10^{10})$ (cf. problem 8). Nevertheless, when we are surprised each time such apparently rare events do occur, the reason is that unconsciously we put the problem differently. If instead of considering a usual deck of cards we had numbered the 52 cards in a random manner from 1 to 52, the result discussed would have corresponded to a certain number of permutations of the numbers from 1 to 52, no one of which we would have found especially conspicuous. Thus it is quite arbitrary which event we make prominent, for unconsciously we put the problem in such a way that we compare only the two possibilities: either 13 cards of the same color or not 13 cards of the same color. However, in that case it is obvious that the probability of the former event is negligible as compared to that of the latter.

Secondly we stress that we cannot demand that the event must not occur in a *series* of observations. For example, if an event has the probability $1/n$, where n is a very large number, the probability of the event occurring at least once among n observations is

$$1 - \left(1 - \frac{1}{n}\right)^n \sim 1 - e^{-1} = 0.63$$

(why?), and this probability is by no means negligible. In fact the event occurs in the mean just once for every n observations (why?).

§9.3. It must be noted that strictly speaking the main criterion for agreement, **A**, cannot be verified directly. For, if an event has a negligibly small probability and if it does not occur in a single observation, the theory apparently agrees with experience. But if the event does occur, nevertheless, is the theory then in agreement with experience or not? However, it follows from the very nature of the problem that a single observation can never prove, or disprove, a *statistical* theory. Proof, or disproof, is possible only by performing several observations, as in the following criteria, I–VI which are deduced from the main criterion **A** and the mathematically proved theorems of

probability, especially those connected with the concept of *convergence in probability* (§ 8.1).

From Bernoulli's theorem (8.1.13) the relative frequency of an event converges in probability to its probability. Thus from this theorem and **A** we must demand as the next criterion for agreement between theory and experience that:

I. *Every event should occur in a series of observations a number of times which is practically speaking proportional to the probability calculated for that event if the number of observations in the series is large.*

In other words, the observed relative frequency should lie very close to the calculated probability. Thus I is, of course, of importance only if the probability of the event has been deduced from other probabilities; otherwise, I is simply equivalent to the definition of the concept of probability (cf. § 1.4).

If for a random variable we have deduced a certain distribution function we must next demand that:

II. *A large number of observed values of a random variable should be distributed according to its theoretical distribution function.*

How such a comparison is performed in practice is discussed in statistics (cf. § 10.4–§ 10.8).

From **A** and (8.1.6) we must demand that:

III. *A function of n observations, the dispersion of which goes to 0 for indefinitely increasing values of n, should practically speaking be equal to its mean value for large values of n.*

In particular, since, from the law of large numbers, the average converges in probability to the mean, we must demand that:

IV. *The average of a large number of independent observations of a random variable should practically speaking be equal to the mean value.*

Thus the mean value indicates the average order of magnitude of the random variable in question.

From **A** and Tschebyscheff's inequality (8.1.4) we must demand that:

V. *An observed value of a random variable should practically speaking not deviate from the mean value by more than a small number times the dispersion.*

Example. In § 8.2 we have extended Bernoulli's theorem to a statement concerning the distribution of a number of relative frequencies about the probability. From (8.2.4) and **A** we must therefore demand that:

VI. *For a large number of series of observations, each containing a large number of observations, the relative frequencies of a certain event should be normally distributed about the probability of the event.*

Finally we stress that we can never ask whether or not the theoretical results agree with the empirical data but only whether they agree sufficiently well (cf. §1.1). Next it follows from the very nature of the problem that it is impossible to give general rules for judging such questions, since this is a subjective question which can be settled only by each person and for each problem separately. However, with respect to this fact, probability does not differ from all other applications of mathematics to the description of the phenomena of the real world, e.g., geometry, astronomy, and theoretical physics.

II. THE PRACTICAL IMPORTANCE OF PROBABILITY

§9.4. As already discussed in Chapter 1 probability is applied whenever phenomena are described and analyzed by means of a statistical description. But after all what is the use of this?

First, probability is used in **statistics** for purely **descriptive** purposes, expressing statistical material in a short and concise way (cf. §10.1). We have important applications of statistics, e.g., in **economics** and **population statistics** as well as in the **theory of errors** and of **adjustment** (cf. Chapters 11 and 12).

§9.5. **Second,** probability is used in **statistics** for purposes of **analysis.** In a multitude of experimental investigations we try to analyze certain phenomena by comparing observations obtained under different conditions. The main principle in any such analysis is to keep as far as possible all factors constant except one which is varied in order to find the influence that it has on the phenomenon considered. However, it is necessary to treat the phenomena considered by means of *statistical methods* as soon as they are random, showing *statistical fluctuations* (§1.1), i.e., that a repetition of the experiment apparently under exactly the same conditions may lead to more or less different results. Thus, if, e.g., in ordnance the chemical composition of the powder is changed and we want to find out whether we can shoot over a longer distance with the new powder, it is, owing to the statistical fluctuations, not enough to fire a single shot with the old and one with the new powder. To get a measure of the statistical fluctuations and thus decide whether the new composition has the result desired we must perform a whole series of shots with the old powder and one series with the new. The problem then is: Is the difference measured **significant,** or is it only what must be expected because of the statistical fluctuations (cf. §11.15)?

Such **statistical analyses** play an ever-increasing role in modern science.[1] Thus, when in **medicine** we want to compare various

[1] A number of such examples may be found in Fisher, *Statistical Methods for Research Workers*, especially Chapter V.

sleeping tablets we must try their effects on a number of patients, and the problem is: Is there any **significant** difference in their effect beyond that which can be expected from the statistical fluctuations? Or in **agricultural science** we may want to compare various fertilizers. We then divide a number of fields in several parts and treat each part with one fertilizer. Again the problem is: Is there any **significant** difference in their effect beyond that which can be expected from the statistical fluctuations? Or in the world of **business** if we want to find out which form for advertising is most effective, we try the various forms of advertising on different groups of people, and the problem is again: Is there any **significant** difference beyond that from the statistical fluctuations? And so forth.

Another form of **statistical analysis** is found in problems of (*a*) **correlation** and (*b*) **regression.** The question here is to investigate (*a*) whether or not certain phenomena depend on each other or are correlated, and (*b*), if so, what the dependency is (cf. §11.18 and §12.15).[1] Such problems are met in **genetics,** e.g., by investigating whether parents of high intelligence are especially likely to have children who are also of high intelligence; or in **agricultural science** by investigating the connection between rainfall and yield; or in **ordnance** by investigating whether there is any connection between deviation of a shot in height and in azimuth (cf. Example, §11.18); and so forth.

A third form of **statistical analysis** is found in **technology** by investigating whether standardized products fulfill certain conditions.[2] Thus, e.g., if a manufacturer wants to make sure that the lifetime of his bulbs is above a certain limit he cannot, of course, examine the lifetime of each bulb produced. Instead he takes samples from the whole production and on the basis of results from tests of such samples he estimates the risk of giving a guarantee for the whole production. It is obvious that such an analysis may have the greatest economic importance.

§9.6. **Third,** probability is used for **predicting** the future course of **random phenomena.** The oldest application of probability for that purpose is to **games of chance.** Another and, in practice, extremely important application is that of **insurance,** in which we want to predict, on the basis of statistical observations regarding death, sickness, fire, theft, and so forth, the number of such events which will occur in the course of a given future time. A third applica-

[1] A number of such examples may be found in Fisher, *Statistical Methods for Research Workers,* especially Chapters V and VI.

[2] Cf., e.g., Fry, *Probability and Its Engineering Uses.*

tion of this kind is found in various branches of **theoretical physics,**[1] where we want to predict the course of experiments concerned with the kinetic theory of gases (cf. the example, §4.8; Exercise 3, §7.4; §4.16), radioactive decay (cf. the example, §3.7; Examples 1 §3.8, 5 §4.3, 2 §4.4) or other phenomena in atomic theory (cf. §4.15, §4.17, Example 6, §5.1), and so forth. Also in **technology,** probability is used for purposes of prediction, e.g., in telephony for designing switchboards when we may ask how many operators a central with given number of subscribers must have in order that the mean waiting time lies below a certain limit (cf. Examples 4 §4.3; 2 §4.4),[2] and so forth. In **ordnance,** probability is used for calculating that fraction of the total number of shots which we should observe in front of the target when the gun is adjusted for maximal number of hits, or for estimating when a shot ought be considered as a stray (cf. §11.17),[3] and so forth. In **business,** probability may be used for estimating how large stocks of a certain commodity a store should have, how many clerks it should have in its various departments (cf. Example 5, §4.3), and so forth.

§9.7. As the above discussion has shown, the application of statistical methods and considerations is a natural procedure in many problems. Thus the role of the statistical description is no exceptional one; on the contrary, it contains the causal description as a special limiting case, since the latter description simply means that each observation gives the same result (cf. p. 3). Because in practice the latter hardly ever occurs in any actual measurements, the statistical description as a rule represents a less idealized description of real phenomena than the causal description does.

Furthermore, the purpose of all science is, first, to state the phenomena in a purely *descriptive* way; second, by *analysis* and by *experiments* to treat apparently scattered and unconnected facts from common points of view, i.e., to put them into order and find their *regularities* and *laws,* to *catalogue* them, so to speak; third, to account quantitatively for the phenomena observed by constructing *theories,* i.e., to lay down rules for calculating in advance quantities which may be compared with observed quantities; and, fourth, to *lead to the discovery of new facts and regularities.* As illustrated above, statistical considerations play a large and ever-increasing role in the pursuit of each of these purposes.

[1] See, e.g., Fürth, *Theoretische Physik.*
[2] See, e.g., Fry, *Probability and Its Engineering Uses.*
[3] Cf., e.g., Hayes, *Elements of Ordnance.*

10.

APPLICATION OF THE THEORY OF PROBABILITY TO STATISTICS

§ 10.1. The purpose of statistics has been formulated most clearly by the English statistician R. A. Fisher:[1]

Statistics *may be characterized briefly as the science of reduction and analysis of observational materials.* As a rule, a statistical material which consists of a certain number of observations $x_1, x_2, \cdots, x_n$ of a random variable x (cf. § 4.1) when given in its "raw form" in which the n numbers are given in the order in which they have been observed is difficult both to survey and to reproduce. Thus it is not suitable for giving us any information about the variable, x, investigated. It is the purpose of statistics to replace the observed material by a relatively few numerical quantities representing the whole material or, in other words, containing as much as possible of the information regarding x as we are looking for.

In statistical material we can seldom include all the observations which might theoretically be performed. Consequently we must, as a rule, regard the given statistical material as a **random sample** which is in itself subject to statistical fluctuations because we would obtain other values, $x_1', x_2', \cdots, x_n'$, if we performed n new observations. *Thus it is, even in the simplest cases, natural to use the concept of probability by introducing certain idealized values, the probabilities, for the relative frequencies of the various possible results of observation* (§ 1.4), *i.e., for the description of the random variable* x *considered to associate with* x *a certain distribution function* $\Phi_x(t)$ (§ 4.2).

As mentioned in § 2.7 it may be a convenient and shorthand expression for this same fact to say that the sample observed has been taken at random from a hypothetical infinite class, the theoretical **population.** This form of language is much used in statistics, although strictly speaking it can be made precise only in the very abstract formulation of probability given by Kolmogoroff (cf. § 2.7).

We note that, as a rule, we are not interested in the actual *empirical* numbers, $x_1, \cdots, x_n$, themselves, but in the *theoretical* concept,

[1] Fisher, *Statistical Methods for Research Workers.*

$\Phi_x(t)$, the *model* by means of which we describe our observations (cf. § 1.1). Just as, e.g., in physics by investigating, say, the law of free fall, $s = \frac{1}{2}gt^2$, we are not interested in the actually measured (s, t)-values, which, owing to the unavoidable measuring errors, will be scattered more or less about this theoretical curve. What we look for here is how to deduce from our observations a numerical value of the parameter g, its "true" value, and, of course, to check whether this theoretical law gives a satisfactory description of our observational material.

§ 10.2. The theoretical distribution function, $\Phi_x(t)$, will, as a rule, contain one or a few constants or *parameters* such as μ and σ in the normal distribution. By stating the numerical values of these parameters the random variable we investigate is completely characterized, or, in other words, our whole observational material is represented by the values of the parameters. Now on the basis of our observations, $x_1, \cdots, x_n$, our task is to **estimate,** or form **estimates** of, the numerical values of the parameters in $\Phi_x(t)$. These empirical estimates of the theoretical parameters are, of course, themselves random variables, being subject to statistical fluctuations. In order to obtain a measure for the expected magnitude of these fluctuations and thus for the certainty with which we can rely on the values for the parameters found from the observations, we must next deduce from $\Phi_x(t)$ the distribution functions of our estimates. Having done that we have solved our problem and have reduced the given observational material as far as possible.

Thus, after the investigation has been planned every statistical problem demands that we answer the following four questions:

I. *The question: which distribution function to associate with the random variable under consideration.*

II. *The question: in what way we can investigate how well the distribution function chosen fits the observations.*

III. *The question: how to calculate from the given sample, i.e., the given observational values $x_1, \cdots, x_n$, the best possible estimates of the unknown parameters in the distribution function.*

IV. *The question: to deduce from $\Phi_x(t)$ the distributions of these estimates and thus to construct methods of testing their uncertainties.*

§ 10.3. Thus, before we can solve a given statistical problem we must first set up a *hypothesis* regarding the mathematical form of the distribution function. Sometimes we know from previous experience that a certain form may be used, e.g., the *normal* (7.1.1). Or we may *deduce* from certain simplifying assumptions about the phenomena considered the distribution function by means of the rules of proba-

bility. As a rule, this is the case in theoretical physics or in the theory of games of chance, as we have seen in many examples in Chapters 1–9.

Example 1. This is also the case in the theory of errors (cf. § 11.4), where we may assume that a great number of independent causes of errors are at work so that the error observed is the sum of all these "elementary errors," as they are called. From the central limit theorem (cf. § 8.3) it follows that the error observed is to a high degree of approximation normally distributed.

Example 2. By a generalization of the "scheme of elementary errors" of the previous example Kapteyn[1] has deduced a whole class of distributions from the normal. Let us assume that our random variable x is not directly a sum of a large number of variables but that its value is due to a large number of causes, giving successively an "impulse," the effect of which depends partly on the magnitude of the impulse and partly on the magnitude given to x by the previous impulses. Let $z_1, z_2, \cdots, z_\nu$ be the independent impulses from ν causes, and let x_i be the result of the effect of the first i of these impulses. Next we assume that x_{i+1} depends only on the value of x_i, but not on the past history, viz., the specific way in which the value of x_i is reached. In other words, we assume that there exists a function $g(x)$ such that

$$x_{i+1} = x_i + z_{i+1} g(x_i), \qquad g(x) \neq 0.$$

Now, if ν is large and, therefore, each contribution is small, we have

$$z_1 + z_2 + \cdots + z_\nu = \sum_{i=0}^{\nu-1} \frac{x_{i+1} - x_i}{g(x_i)} \sim \int_{x_0}^{x} \frac{dx}{g(x)} = G(x). \tag{1}$$

Under very general conditions $z_1 + \cdots + z_\nu$ will be approximately normally distributed for large ν, by the central limit theorem (§ 8.3). Thus in this case we have that not x, itself, but a certain function, $G(x)$, given in (1) is normally distributed: Kapteyn's class of distributions which we have already considered in (7.7.1):

$$d\Phi_x(t) = d\Psi\left(\frac{G(t) - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(G(t) - \mu)^2}{2\sigma^2}\right] \left|\frac{dG(t)}{dt}\right| dt$$

$$\mathfrak{M}\{G(x)\} = \mu, \qquad \sigma\{G(x)\} = \sigma. \tag{2}$$

In many sociological and biological problems Kapteyn's scheme may be applied, if we put $g(x) = x$, i.e., if we assume that the effect of a

[1] Kapteyn and van Uven, *Skew Frequency Curves.*

given cause is proportional to the magnitude already reached. Since $\int \frac{dx}{x} = \ln x$, we find the distribution, called the **logarithmico-normal**:

$$d\Phi_x(t) = d\Psi\left(\frac{\ln t - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(\ln t - \mu)^2}{2\sigma^2}\right] \frac{dt}{t} \tag{3}$$

$$(0 \leqq t < \infty)$$

$$\mathfrak{M}\{\ln x\} = \mu, \qquad \sigma\{\ln x\} = \sigma.$$

For example, this formula gives a satisfactory description of the distribution of the weight of school children, of income within certain vocations; and so forth.

In general we cannot deduce the distribution function theoretically but must try to obtain it directly from the observations. Now, since we can always perform only a finite number of observations, a multitude of theoretical formulae may be fitted to a given observational material, no one of which is either "true" or "false" (cf. § 1.1). But they may fit the observations more or less satisfactorily, and those giving only a poor fit may be excluded by various tests (cf. § 10.4). In general, we must choose a mathematical form for $\Phi_x(t)$ which we either know analytically or which has been numerically tabulated.[1] In modern statistics, Kaptcyn's class of distributions, given in (2), is used more and more frequently because in this case the distribution of the estimates may be obtained. Also, however, some of the older classes of distributions are still in use: for example, the class of Pearson obtained as solutions of the differential equation

$$\frac{d\varphi(t)}{dt} = \frac{(t - \alpha)\varphi(t)}{\beta_0 + \beta_1 t + \beta_2 t^2}, \tag{4}$$

where α, β_0, β_1, and β_2 are four parameters; or the class of Gram-Charlier which consists of series expansions after the normal distribution, $\Psi\left(\frac{t - \mu}{\sigma}\right)$, and its derivatives.[1]

§ 10.4. We now assume that question I has been answered, a hypothesis regarding $\Phi_x(t)$ having been laid down. Next we have to answer question II, i.e., to work out methods to test how well the theoretical distribution fits the observations. The simplest test consists of plotting the observations graphically and drawing the corresponding theoretical curve for comparison.

[1] For general distributions see, e.g., Kendall, *The Advanced Theory of Statistics*.

Let us first treat the case of *discontinuous distributions* (§ 4.3). We plot the possible values, t_i, as abscissae and as ordinates the relative frequencies

$$f_i = \frac{n_i}{n}, \tag{1}$$

where n_i is the number among the n observations in which the result t_i has been observed. This graph is said to give the **empirical, observed, frequency** or **sample distribution.** For the comparison we plot in the same graph the **theoretical** or **true** distribution, viz., the probabilities φ_i for the results t_i. In spite of the fact that φ_i is defined only for $t = t_i$, it may be convenient for the comparison to connect the *theoretical* points with a smooth curve, as shown, e.g., in Fig. 7. Also we may plot as ordinates the absolute frequencies, n_i, which, if n_i are only small numbers, is done most conveniently by plotting above each t_i a number of dots equal to the number n_i, as, e.g., shown in Fig. 1, p. 7. For comparison we plot the theoretical values of n_i, viz., $\nu_i = n\varphi_i$.

Since f_i is the relative frequency among the n independent observations of the event $x = t_i$ with probability φ_i, f_i is for each fixed value of i a binomially distributed variable for which from (6.4.14) and (6.4.15)

$$\mathfrak{M}\{f_i\} = \varphi_i \tag{2}$$

$$\sigma^2\{f_i\} = \frac{\varphi_i(1-\varphi_i)}{n} \xrightarrow[n\to\infty]{} 0. \tag{3}$$

Thus from Bernoulli's theorem (8.1.13) $f_i \xrightarrow[n\to\infty]{\text{in } p} \varphi_i$; i.e., *we must, for each fixed value of* t_i, *expect the observed value of* f_i *to lie very close to* φ_i *for large values of* n.

Example. Let the random variable be the number x of radioactive atoms decaying in a certain interval of time. From Example 5, § 4.3, x has Poisson's distribution. By a measurement of, in all, $n = 2608$ time intervals (each of length $\frac{1}{8}$ minute) the values given in Table 1 were found.[1] Here n_i denotes the number of time intervals in which i atoms were found decaying and $\nu_i = n\varphi_i$ the corresponding theoretical number, φ_i being calculated from (4.3.5) with $\mu = 3.87$ (cf. Example 3, § 10.10). In Fig. 7, n_i and ν_i are plotted against i. It is seen from both the table and the figure that the agreement between

[1] See Rutherford, Chadwick, and Ellis, *Radiations from Radioactive Substances*, p. 172, London, 1930.

TABLE 1

Number of Decays in One Interval	Observed Number of Intervals	Theoretical Number of Intervals
i	n_i	ν_i
0	57	54
1	203	210
2	383	407
3	525	525
4	532	508
5	408	394
6	273	255
7	139	140
8	45	68
9	27	29
10	10	11
11	4	4
12	0	1
13	1	1
14	1	1
	$n = \sum_i n_i = 2608$	$\nu = \sum_i \nu_i = 2608$

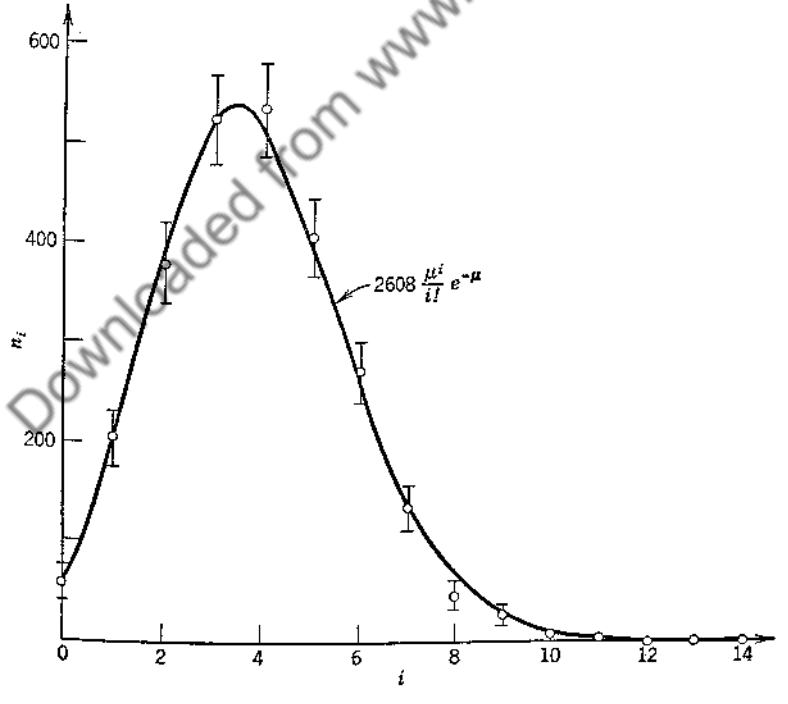

FIG. 7.

theory and experiment, i.e., between the theoretical and the sample distribution, is very satisfactory (cf. Example 1, § 10.8).

§ 10.5. Let us now consider *continuous distributions* (§ 4.4). We perform a **grouping** of the observations, which means that we divide a suitable interval which comprises all the observed values $x_1, x_2, \cdots, x_n$ in a certain number $m < n$ of subintervals with lengths Δt_i, called the **class intervals,** all of which need not have the same lengths. The middle points of these intervals we denote $t_1, t_2, \cdots, t_m$. For each t_i we count the number of observations, n_i, for which the results lie in the ith class interval, i.e., for which

$$t_i - \frac{\Delta t_i}{2} < x \leqq t_i + \frac{\Delta t_i}{2}. \tag{1}$$

Next we form the corresponding relative frequencies divided by Δt_i

$$f(t_i) = \frac{1}{\Delta t_i} \frac{n_i}{n} \tag{2}$$

and plot in a tu-coordinate system the step-function

$$u = f(t) = f(t_i) \qquad \text{for} \qquad t_i - \frac{\Delta t_i}{2} < t \leqq t_i + \frac{\Delta t_i}{2}. \tag{3}$$

This graph is called a **histogram.** We see that the relative number of observations lying in a certain interval is equal to the *area* between the curve $u = f(t)$ and the interval on the t-axis. This graph of $u = f(t)$ gives a convenient representation of the observations if n is not too small and the Δt_i's are chosen in a suitable way (cf. below). It is said to give the **empirical, observed, frequency** or **sample distribution.** For comparison we next plot in the same graph the corresponding **theoretical,** or **true, distribution,** viz., the probability density $u = \varphi(t)$. Also it may be convenient to plot as ordinates the absolute frequencies $n_i/\Delta t_i$ and then for comparison plot the corresponding theoretical curve $u = n\varphi(t)$. If the class intervals are all of equal length, Δt, it may be convenient to omit the division by Δt and plot n_i and for comparison $n\,\Delta t\varphi(t)$.

Sometimes we plot the points $(t_i, f(t_i))$, $\left(t_i, \frac{n_i}{\Delta t_i}\right)$ or (t_i, n_i) and then connect these points by straight lines, thus obtaining a smoother curve, called the **frequency polygon.** However, this is less appropriate than the histogram since by doing this we veil the fact that, in contrast to a theoretical distribution, every empirical distribution is dis-

continuous, it being possible to make only a finite number of observations, the results of which are all integers when expressed in the smallest unit possible with our measuring instrument.

In analogy with (10.4.2) and (10.4.3) $f(t_i)$ is, for each fixed value of i, a binomially distributed variable with

$$\mathfrak{M}\{f(t_i)\} = \frac{1}{\Delta t_i}\,\mathfrak{M}\left\{\frac{n_i}{n}\right\} = \frac{1}{\Delta t_i}\int_{t_i-(\Delta t_i/2)}^{t_i+(\Delta t_i/2)} \varphi(t)\,dt = \varphi(\xi_i)$$

$$t_i - \frac{\Delta t_i}{2} < \xi_i < t_i + \frac{\Delta t_i}{2} \tag{4}$$

$$\sigma^2\{f(t_i)\} = \frac{1}{(\Delta t_i)^2}\,\sigma^2\left\{\frac{n_i}{n}\right\} = \frac{\varphi(\xi_i)(1-\varphi(\xi_i)\,\Delta t_i)}{n\,\Delta t_i}. \tag{5}$$

Thus from Bernoulli's theorem (8.1.13) $f(t_i) \xrightarrow[n\to\infty]{\text{in } p} \varphi(\xi_i)$; *i.e., we must, for each fixed value of* t_i, *expect the observed value of* $f(t_i)$ *to lie very close to* $\varphi(\xi_i)$ *for large values of* n, $\varphi(\xi_i)$ being the average value of $\varphi(t)$ in the class interval Δt_i. For small values of Δt_i, $\varphi(\xi_i)$ is practically equal to $\varphi(t_i)$.

If we plot n_i we have to calculate the corresponding theoretical values

$$\nu_i = n\int_{t_i-(\Delta t_i/2)}^{t_i+(\Delta t_i/2)} \varphi(t)\,dt \tag{6}$$

for each value of i and compare n_i with ν_i, which is most conveniently done in a table (cf. example).

In order to get as detailed a comparison as possible we see that the class intervals must be chosen as small as possible. However, from (5) we see that, the smaller the Δt_i's, the larger the statistical fluctuations of $f(t_i)$, which results in the histogram's assuming a more irregular shape (cf. Figs. 10 and 11). Thus n must be chosen in such a way that the Δt_i's can be chosen small and at the same time $n\,\Delta t_i$ are large numbers. *In practice we usually try to choose* n *and the* Δt_i*'s so that each class interval contains at least* 5 *observations* (except possibly the extreme class intervals). If the Δt_i's are too small the statistical fluctuations will dominate, and if they are too large the details of the distribution will be quenched. Thus except when n is very large the histogram should be used only for a rough representation, while the sum polygon (§ 10.6) should be used for all detailed comparisons with the theoretical distribution.

Example. By shooting 96 shots with an 8-mm machine gun against a target at a distance 300 m the tabulated deviations in azimuth and height from the target having coordinates (0, 0) have been measured.

TABLE 1

Azimuth Deviation in Centimeters		
t_i	n_i	ν_i
−30	0	0.17
−20	2	1.80
−10	9	9.73
0	28	25.43
10	30	32.22
20	21	19.83
30	5	5.91
40	1	0.85
50	0	0.06
	$\sum_i n_i = n = 96$	$\sum_i \nu_i = \nu = 96.00$

TABLE 2

Height Deviation in Centimeters		
t_i	n_i	ν_i
−60	0	0.54
−50	3	1.95
−40	5	5.92
−30	13	12.86
−20	18	19.94
−10	21	22.07
0	21	17.46
10	10	9.86
20	5	3.98
30	0	1.42
	$\sum_i n_i = n = 96$	$\sum_i \nu_i = \nu = 96.00$

Here n_i is the number of shots for which the deviation in azimuth, or height, lies in the class about t_i with $\Delta t = 10$ cm, and ν_i are given by (6) and the normal distribution, i.e.,

$$\nu_i = 96 \int_{t_i-(\Delta t/2)}^{t_i+(\Delta t/2)} \frac{1}{\sigma}\psi\left(\frac{t-\mu}{\sigma}\right) dt =$$

$$96\left(\Psi\left(\frac{t_i+\frac{\Delta t}{2}-\mu}{\sigma}\right) - \Psi\left(\frac{t_i-\frac{\Delta t}{2}-\mu}{\sigma}\right)\right), \quad (7)$$

where Ψ is tabulated in Table I and the parameters have the numerical values $\mu = 8.3$ cm and $\sigma = 11.4$ cm for the azimuth and $\mu = -12.0$ cm and $\sigma = 17.0$ cm for the height (cf. example, §11.12).

In Figs. 8 and 9 we have plotted the two histograms for the deviations of azimuth and height, respectively. (We have omitted the division with $n = 96$.) For comparison we have plotted the corresponding theoretical curves $96\,\frac{1}{\sigma}\psi\left(\frac{t-\mu}{\sigma}\right)$ obtained from Table I

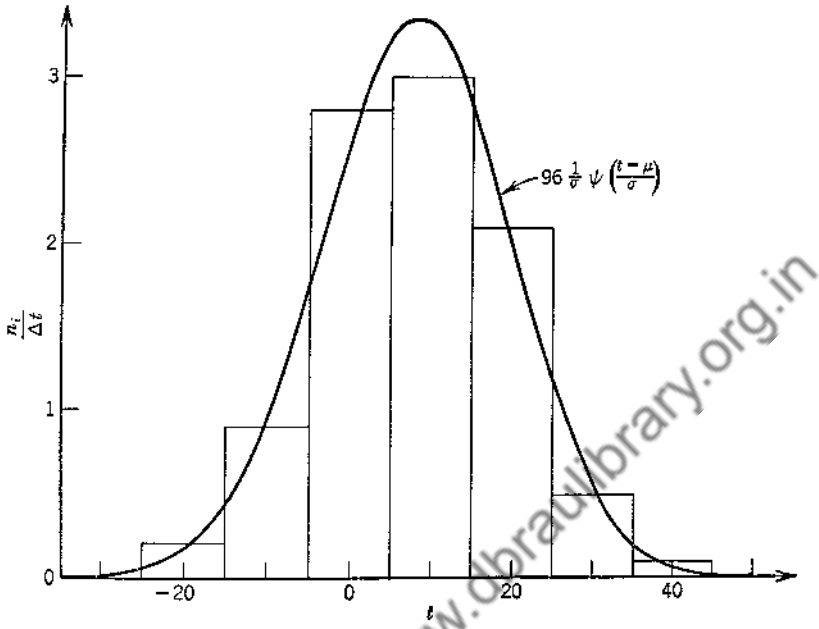

FIG. 8. Azimuth deviation. $\Delta t = 10$ cm.

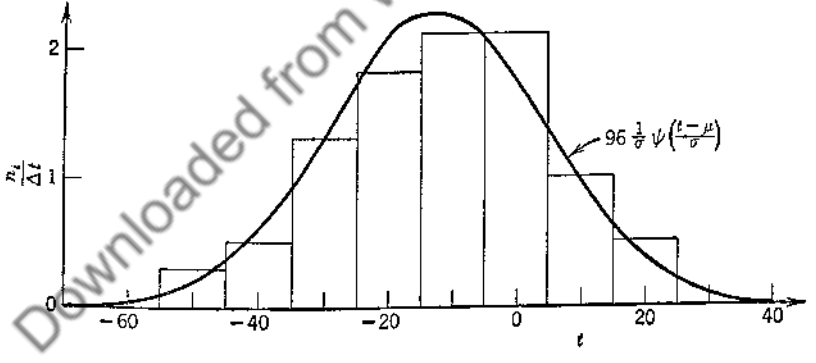

FIG. 9. Height deviation. $\Delta t = 10$ cm.

with the values of μ and σ mentioned previously. It is seen from both the tables and the figures that the agreement between observations and theory is very satisfactory. As a matter of fact we shall see in the next two topics that the agreement is even better than Figs. 8 and 9 show.

In Figs. 10 and 11 we have plotted the histograms of the same observations for a class interval $\Delta t = 2$ cm. The figures clearly show that

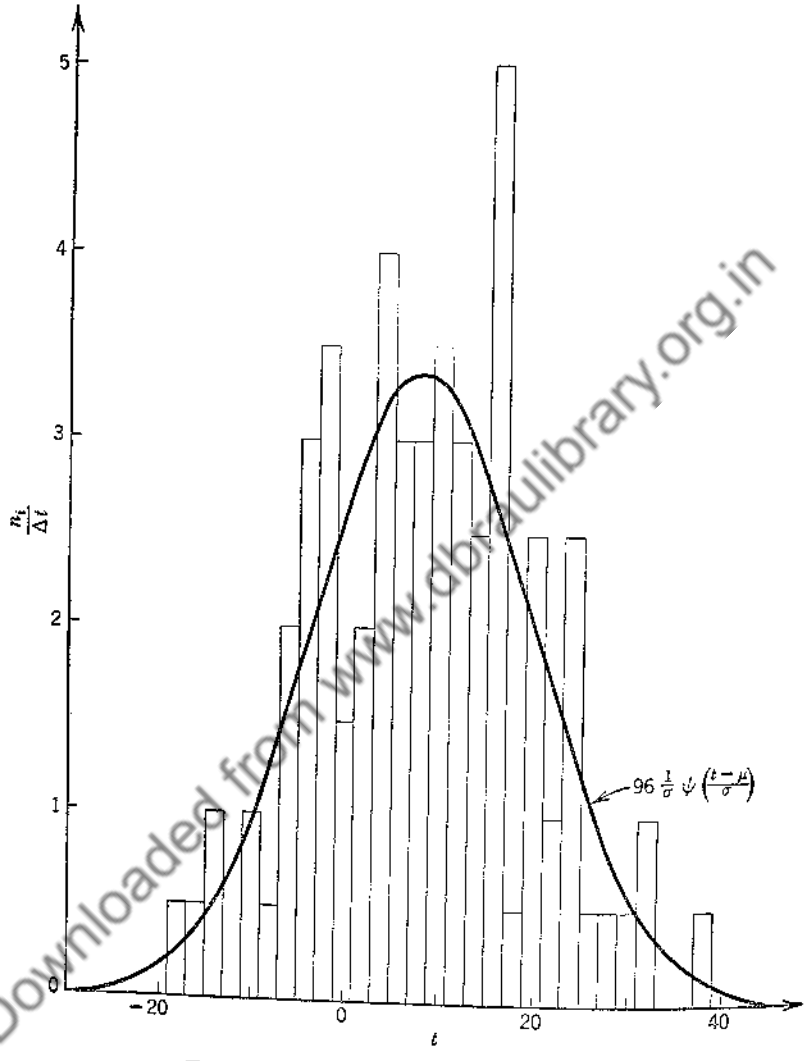

FIG. 10. Azimuth deviation. $\Delta t = 2$ cm.

by this the statistical fluctuations are increased so strongly that it is highly doubtful, from these curves alone, whether the observations agree with the theoretical curves or not.

§ 10.6. A method that may always be applied to both discontinuous and continuous distributions consists in plotting the **sum polygon**

$$u = F(t) = \frac{N(t)}{n}, \tag{1}$$

where $N(t)$ is the number of observations for which the result is smaller than or equal to t. $F(t)$ is also called a **cumulated histogram** or **empirical, observed, frequency,** or **sample distribution function.** For comparison we plot in the same graph the corresponding **theoretical** or **true distribution,** viz., the distribution function

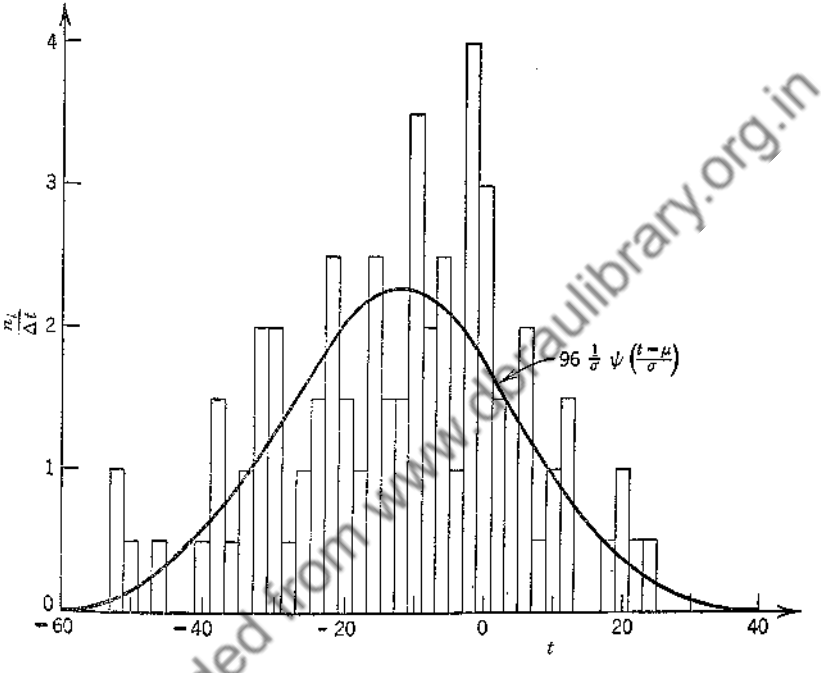

FIG. 11. Height deviation. $\Delta t = 2$ cm.

$u = \Phi(t)$. It may also be convenient here to omit the division by n in (1) and plot the corresponding theoretical curve $\nu = n\Phi(t)$. Since

$$\mathfrak{M}\{\boldsymbol{F}(t)\} = \Phi(t) \tag{2}$$

$$\sigma^2\{\boldsymbol{F}(t)\} = \frac{\Phi(t)(1 - \Phi(t))}{n} \tag{3}$$

(check) *we have from Bernoulli's theorem* (8.1.13) *that* $\boldsymbol{F}(t) \xrightarrow[n\to\infty]{\text{in } p} \Phi(t)$; *i.e., we must, for each fixed value of t, expect the observed value of* $F(t)$ *to lie very close to* $\Phi(t)$ *for large values of n.* This method gives the most detailed comparison between observation and theory. However, for large values of n it is cumbersome to plot the sum polygon, but in

such cases we may group the observations with a small class interval Δt before calculating the sum polygon.

Example. For the deviations in azimuth and height of the 96 shots in the example, § 10.5, we obtain the two sum polygons shown in Figs 12 and 13 (the division by n having again been omitted). For comparison we have plotted the corresponding theoretical curves, viz.,

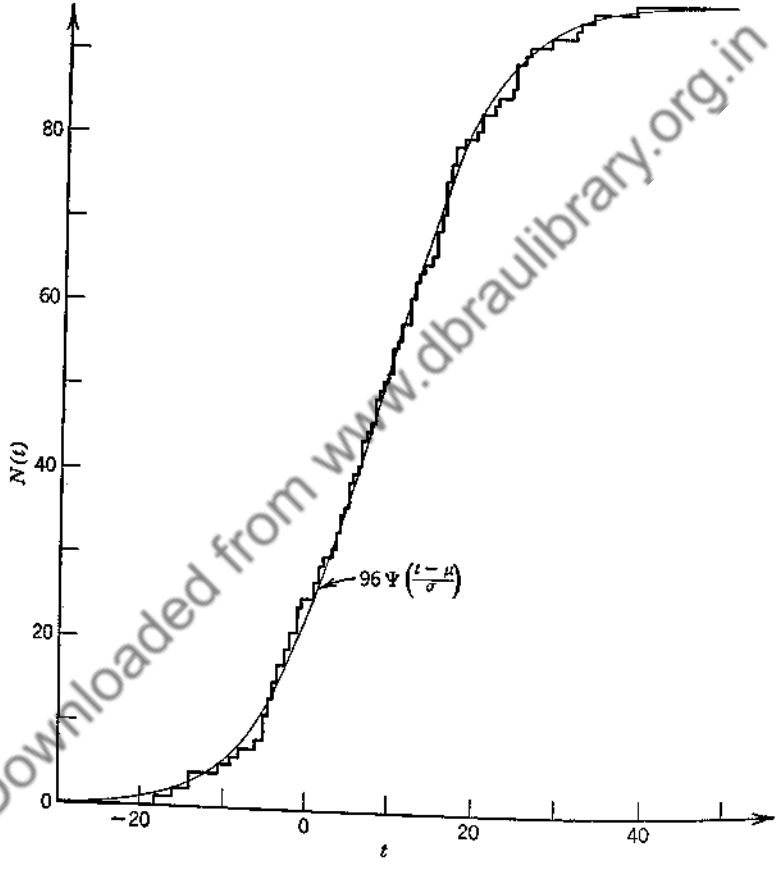

FIG. 12. Azimuth deviation.

$96\Psi\left(\frac{t-\mu}{\sigma}\right)$, with the previously stated numerical values of μ and σ. It is seen that the agreement is extremely satisfactory and much better than is shown by the previous figures, especially Figs. 10 and 11.

§ 10.7. For Kapteyn's distributions, given in (10.3.2), which are generalizations of the normal distribution, we have an especially

simple method for comparing observations with theory. Since $u = \Psi(t)$ is always increasing in the whole interval $-\infty < t < \infty$ *we may give* t *as a function of* u, $t = \Psi^{-1}(u)$.[1] For the general normal distribution, $u = \Psi\left(\frac{t-\mu}{\sigma}\right)$, we thus have $\Psi^{-1}(u) = (t-\mu)/\sigma = v_1(t)$. Since $F(t)$ is expected to lie close to u, we shall expect the function

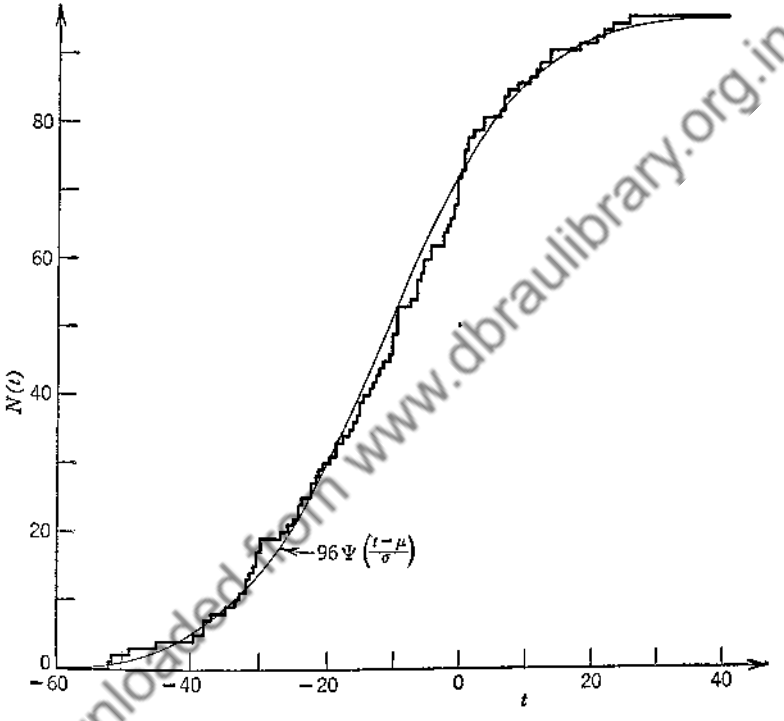

FIG. 13. Height deviation.

$v_2 = \Psi^{-1}(F(t))$, $F(t)$ given in (10.6.1), to lie very close to the straight line $v_1 = (t-\mu)/\sigma$ passing through the point $(\mu, 0)$ and having the slope $1/\sigma$. This method, sometimes called the **probit diagram** method, is very convenient. Since the normal distribution has been tabulated only for the values $\mu = 0$ and $\sigma = 1$ it saves one a lot of computing. (There also exists so-called probability paper for performing this transformation graphically.)

If not x, but a certain function of x, $G(x)$, is normally distributed we shall expect $v_2 = \Psi^{-1}(F(t))$ to lie very close to the curve $v_1 =$

[1] This function is tabulated in Fisher and Yates, *Statistical Tables*, Table IX.

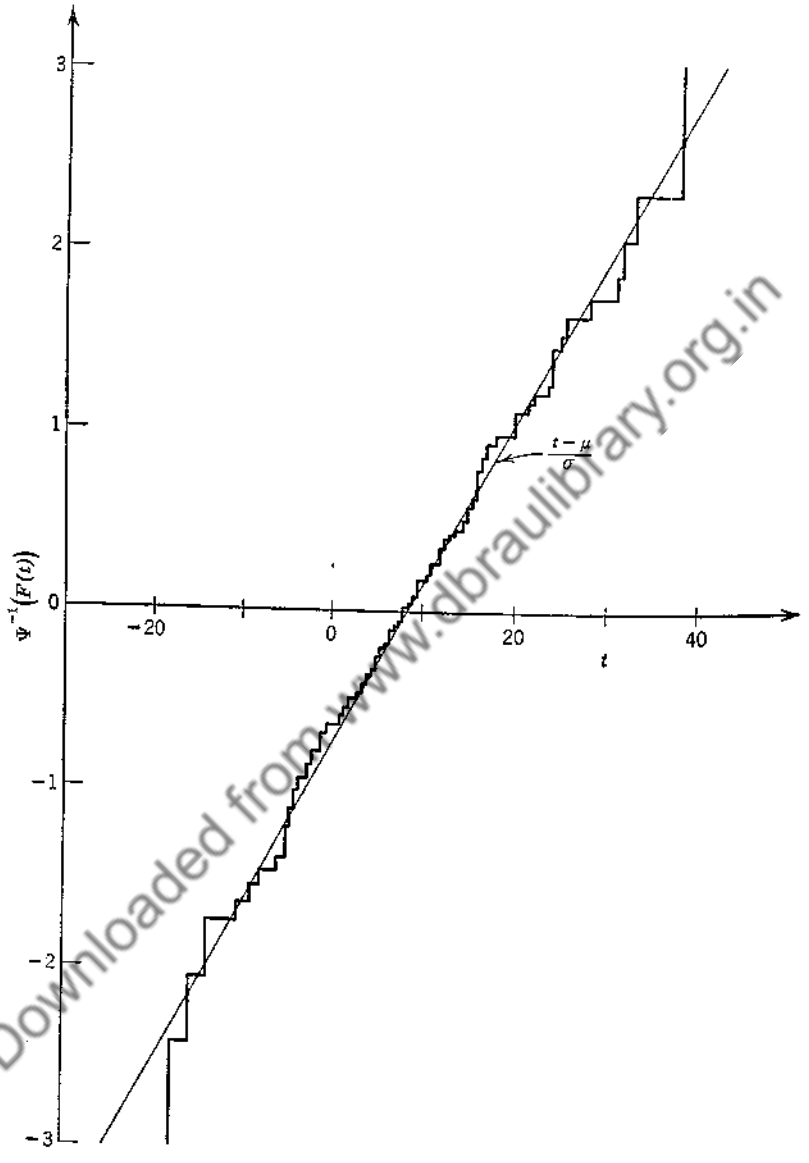

FIG. 14. Azimuth deviation.

$\frac{G(t)-\mu}{\sigma}$. Thus the probit diagram immediately indicates which function of x to choose. For the normal distribution itself this curve is, as mentioned, a straight line which is especially convenient since the human eye is very sensitive to even small deviations from a straight

line. Another great advantage of the probit diagram in this last case is that without knowing the numerical values of the parameters it permits testing the normal *form* of the distribution function by only seeing whether or not the curve is approximately a straight line.

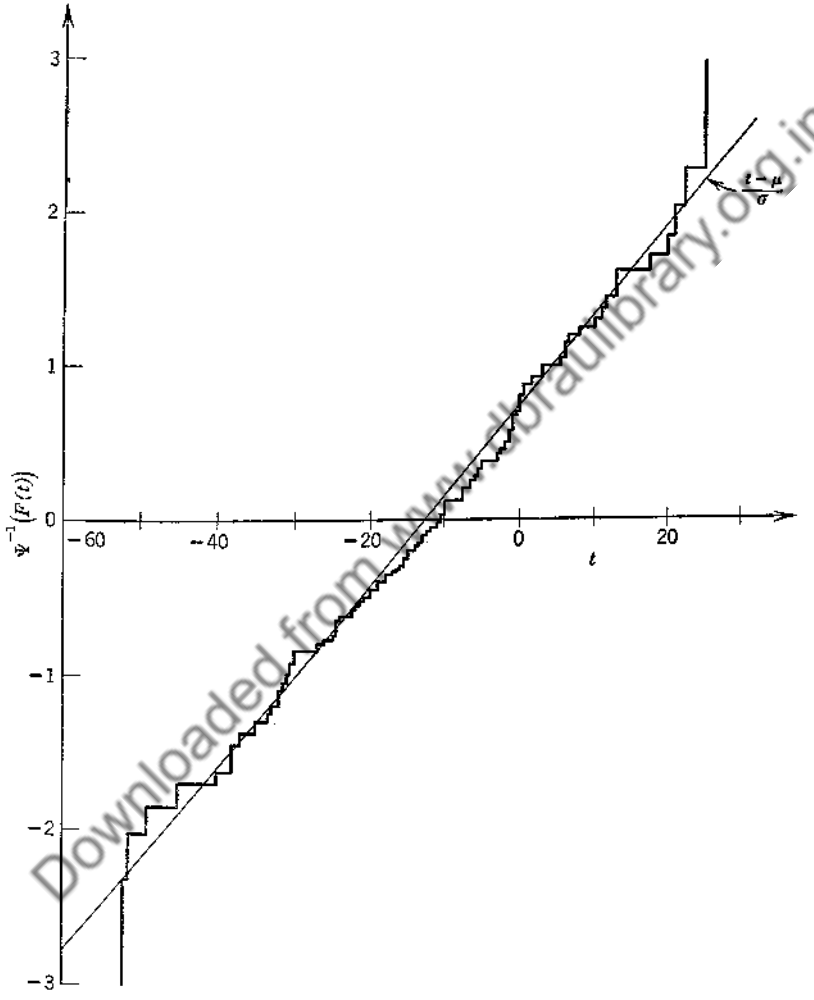

FIG. 15. Height deviation.

Furthermore, it may be used for obtaining the numerical values of the parameters μ and σ graphically.

Example 1. In Figs. 14 and 15 we have plotted the probit diagrams for the deviation in azimuth and height of the 96 shots of the

example, § 10.5, and for comparison the corresponding straight lines $v = (t - \mu)/\sigma$. Again it is seen that the agreement is extremely satisfactory and much better than shown by the Figs. 8 and 9 or 10 and 11.

Exercise. We note that Figs. 14 and 15 show that the statistical fluctuations are larger at the ends than around $t = \mu$. Show that this is just what must be expected because

$$\sigma^2\{\Psi^{-1}(F(t))\} \sim \frac{1}{\psi^2(v)} \frac{\Psi(v)(1 - \Psi(v))}{n} \underset{|v| \to \infty}{\to} \infty, \qquad v = \frac{t - \mu}{\sigma}. \tag{1}$$

Example 2. For 750 electricity consumers the distribution with regard to consumption, measured in hours of use, t, corresponding to

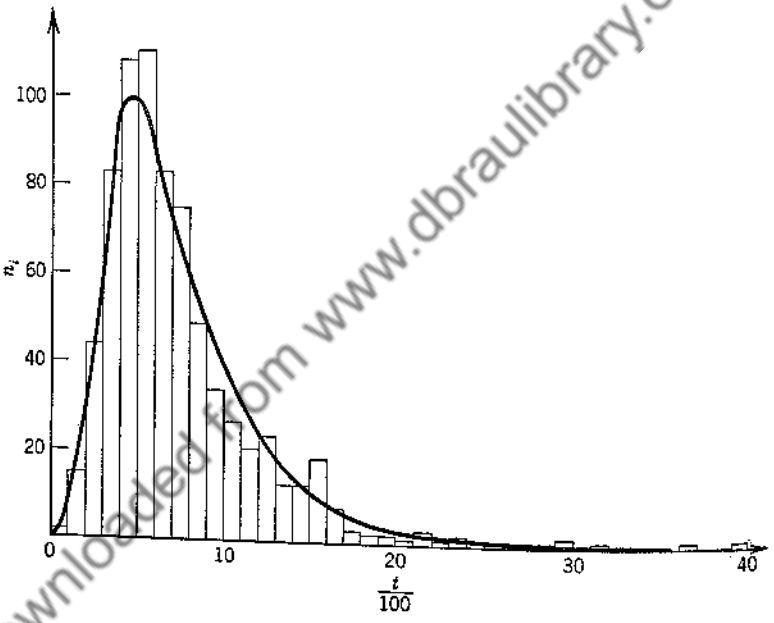

FIG. 16.

maximum load, is shown by the histogram in Fig. 16 ($\Delta t = 100$ hours). The corresponding probit diagram is shown in Fig. 17 and obviously does not give a straight line, but a logarithm-like curve. In fact, if we plot the probit diagram against $\ln t$, we get a straight line as shown in Fig. 18. Reading off the values of μ and σ from this graph and calculating $n\,\Delta t\varphi(t)$ by means of (10.3.3) we get the theoretical curve drawn on Fig. 16. Both from Fig. 16 and from Fig. 18 we see that the distribution in question is logarithmico-normal.

§ 10.8. From a graphical picture of a histogram and the corresponding theoretical curve, or f_i and φ_i for a discontinuous distribu-

tion, it is possible to judge the agreement between observations and theory. Any large disagreement is very conspicuous, but it is difficult from such a picture alone to judge in a quantitative way whether the observations deviate more from the theoretical curve than must be expected from the statistical fluctuations.[1] Now, from Laplace's formula (8.2.1) the various frequencies n_i and $N(t)$ are for each fixed i and t approximately normally distributed for large values of n.

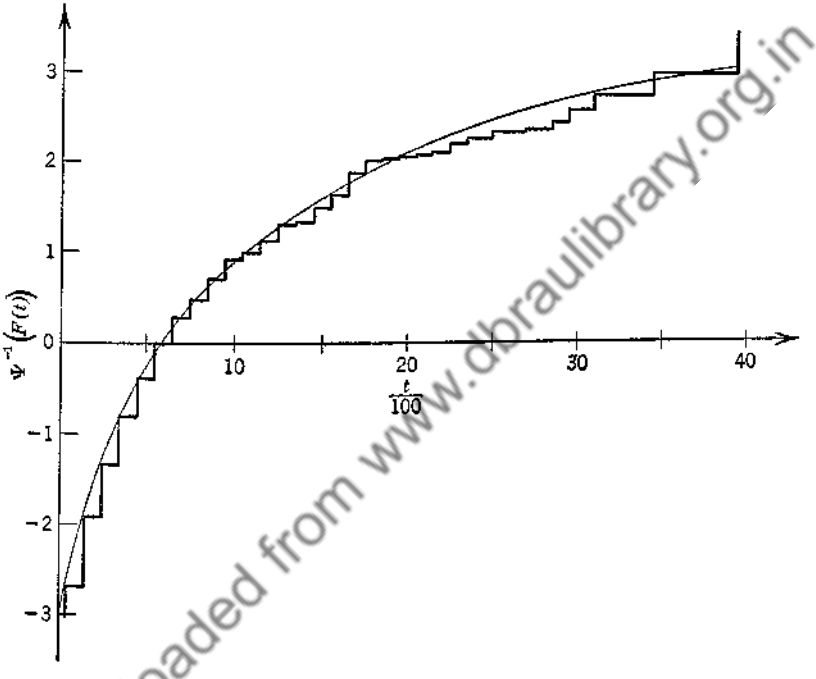

FIG. 17.

From the dispersions (10.4.3), (10.5.5), and (10.6.3), respectively, we can calculate by means of Table II the tolerance limits belonging to any probability P and then see whether the frequencies observed lie within these limits. In statistical practice it is common to take the 5% limit which from (7.4.4) is very nearly the interval $(\mu - 2\sigma, \mu + 2\sigma)$. These tolerance limits may be plotted directly on the graphs themselves and then they give a good idea of the extent of agreement. We note that the choice of the value 5% is, of course, quite arbitrary

[1] As a matter of fact, since it is rather unlikely that all the points lie exactly on the theoretical curve, it is sometimes argued that too good an agreement also shows that the theoretical curve is not the appropriate description of the observations.

and is due only to the fact that it has been found a convenient value in practice. As stressed in § 9.3 it is always a question of personal choice where to put the distinction between satisfactory and unsatisfactory agreement.

As a rule, φ_i, $\varphi(\xi_i)\,\Delta t_i$, and $\Phi(t)$ are small numbers, and since they occur only under a square root we may for a rough estimate of the

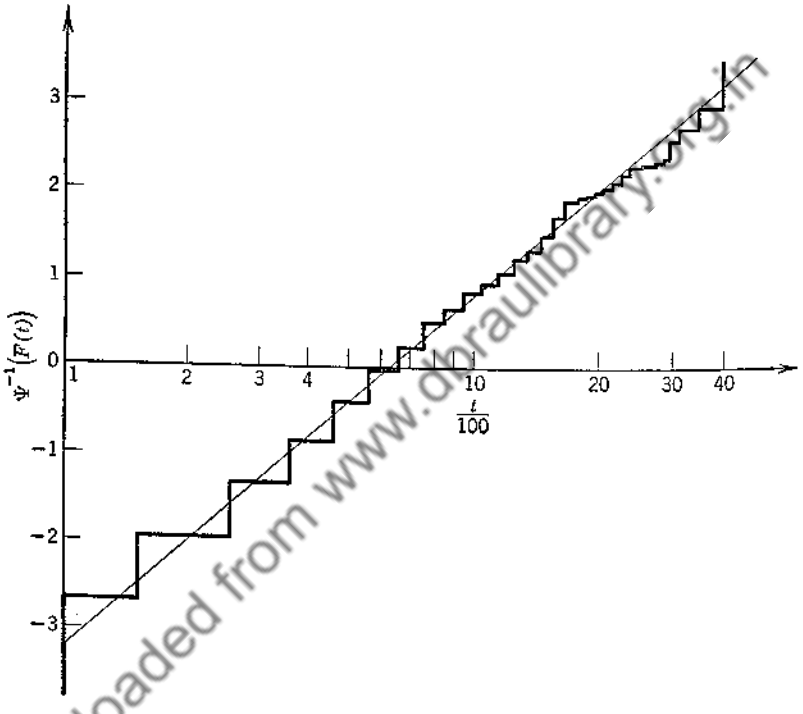

FIG. 18.

tolerance limits write, respectively,

$$\sigma\{n_i\} = \sqrt{n\varphi_i}\,\sqrt{1-\varphi_i} = \sqrt{\mathfrak{M}\{n_i\}}\,\sqrt{1-\varphi_i} \sim \sqrt{n_i}, \tag{1}$$

$$\sigma\{n_i\} = \sqrt{n\varphi(\xi_i)\,\Delta t_i}\,\sqrt{1-\varphi(\xi_i)\,\Delta t_i} = \sqrt{\mathfrak{M}\{n_i\}}\,\sqrt{1-\varphi(\xi_i)\,\Delta t_i} \sim \sqrt{n_i}, \tag{2}$$

$$\sigma\{N(t)\} = \sqrt{n\Phi(t)}\,\sqrt{1-\Phi(t)} = \sqrt{\mathfrak{M}\{N(t)\}}\,\sqrt{1-\Phi(t)} \sim \sqrt{N(t)}, \tag{3}$$

assuming that the actually observed values deviate but slightly from the mean values. This means that we may estimate the dispersions

directly from the observations, which is very convenient, especially if the theoretical distributions are not known beforehand.

Example 1. In Fig. 7 the vertical lines denote the intervals $(n_i - 2\sqrt{n_i},\ n_i + 2\sqrt{n_i})$. It is seen that the theoretical curve lies within these limits.

Example 2. Often the results are given in the form

$$n_i \pm \sqrt{n_i}, \tag{4}$$

but we stress that this is very misleading, since it may induce the belief that *no* values at all may be found outside this interval, although this interval is only an estimate of the tolerance limits belonging to one definite probability, being moreover quite arbitrary, viz., roughly $\frac{2}{3}$ (cf. (7.4.3)).

Example 3. Instead of the 5% limits, being approximately $2\sqrt{n_i}$, some writers also give, mostly because of an old tradition, the "probable" limits, i.e., the 50% limits, which are approximately $\frac{2}{3}\sqrt{n_i}$. However, this is a highly inappropriate procedure, since it underestimates the allowed statistical fluctuations and thus makes the observations appear much more accurate than they are. This may be dangerous, e.g., if the observations are used to distinguish between two theoretical distributions.

In the methods discussed so far each relative frequency was treated independently. Obviously the whole empirical distribution is a random variable, although a many-dimensional one. Other methods for judging quantitatively the extent of agreement have therefore been worked out by treating the empirical distribution as a whole (cf. § 12.16). However, it is beyond the scope of this book to discuss in detail these methods, by which we may obtain objective criteria for rejecting theoretical distributions which disagree with the observations so much as to make them unlikely.[1]

§ 10.9. We shall now treat the third question of § 10.2, that of estimating the parameters. This problem is the main problem of mathematical statistics. Let us consider a random variable x, one- or many-dimensional, and let us assume that, hypothetically, we have associated with x a distribution function $\Phi_x(x)$ of a definite mathematical form but containing one or more parameters, $\theta_1, \cdots, \theta_k$.[2] The

[1] See, e.g., Fisher, *Statistical Methods for Research Workers*, Chapter IV; Cramér, *Mathematical Methods of Statistics*, Chapter 30.

[2] For convenience we shall use throughout the rest of the book the same letter for a random variable, $x, \cdots$, and for the variable in the distribution function, $x, \cdots$. (For many-dimensional random variables the letters x, and so on,

problem then is to estimate from a sample consisting of a certain number, n, of observations, $x_1, x_2, \cdots, x_n$, the "best" numerical values of $\theta_1, \cdots, \theta_k$. First of all it is obvious that any estimate, t, for one of the parameters, θ, will be a certain function of $x_1, \cdots, x_n$, $t = t(x_1, \cdots, x_n)$, but not of $\theta_1, \cdots, \theta_k$. Thus t itself is a random variable since a new sample will give a new value of t because of the statistical fluctuations. Such a function is called a **statistic,**[1] and we write

$$\theta \approx t, \tag{1}$$

read: θ has the estimate t. The distribution function of x being given, that of any statistic is uniquely given and will, as a rule, again contain the parameters $\theta_1, \cdots, \theta_k$. Now the ideal estimate of one of the parameters θ, would, of course, be such a statistic $\boldsymbol{t}$ that its distribution function is for all values of n the causal distribution $\epsilon(t - \theta)$ given in (4.3.3). For in that case each measurement of $\boldsymbol{t}$ would with certainty give as a result the value of θ for which we are looking. However, such an ideal situation hardly ever arises in practice for any *finite* value of n, but only in the limiting case of $n \to \infty$. Therefore it is natural to demand that a statistic, $\boldsymbol{t}$, can be regarded only as a suitable estimate of θ if $\boldsymbol{t} \xrightarrow[n\to\infty]{\text{in } p} \theta$. Such statistics are called **consistent** estimates; from **A,** § 9.2, it follows that *an observed value of* $\boldsymbol{t}$ *will then practically speaking be equal to* θ *for large values of* n. From Tschebyscheff's inequality (8.1.4) it follows that a sufficient (but not necessary) condition for $\boldsymbol{t}$ being a consistent estimate of θ is that $\mathfrak{M}\{\boldsymbol{t}\} \xrightarrow[n\to\infty]{} \theta$ and $\sigma^2\{\boldsymbol{t}\} \xrightarrow[n\to\infty]{} 0$. In particular, if, for all n, $\mathfrak{M}\{\boldsymbol{t}\} = \theta$, $\boldsymbol{t}$ is said to be an **unbiased** estimate of θ.

On the other hand consistency does not tell us anything about the properties of $\boldsymbol{t}$ for the relatively small values of n usually met in practice. Furthermore, there exist infinitely many consistent estimates for one and the same parameter θ and distribution $\Phi_x(x)$: e.g., if $\boldsymbol{t}$ is a consistent estimate of θ, then obviously $g_n\boldsymbol{t}$ is another if g_n is a quantity such that $g_n \xrightarrow[n\to\infty]{} 1$. In order to avoid this trivial possibility it is natural

---

are vectors, standing for $(x_1, x_2, \cdots)$ and so on; for simplicity we shall consider in this chapter only one-dimensional random variables, but the generalization to many-dimensional variables is straightforward.) Furthermore we shall denote observed values as $x_1, x_2, \cdots$. We also remind the reader that we denote all theoretical concepts by Greek letters and corresponding empirical quantities by the corresponding Roman letters, e.g., θ and t.

[1] Most of the terminology used here has been introduced by R. A. Fisher. See his papers: "Mathematical Foundations of Theoretical Statistics" and "Theory of Statistical Estimation."

to "normalize" the estimates by demanding that they be unbiased, which may always be done if $\mathfrak{M}\{t\}$ exists. Moreover, it is natural to consider one consistent, unbiased statistic t_1 a better estimate of θ if its probability mass is more concentrated around θ than that of another consistent, unbiased estimate t_2, the concentration being measured by any suitable measure of dispersion. As such we shall consider here only the dispersion, σ, given by $\sigma^2\{t\} = \mathfrak{M}\{(t - \mathfrak{M}\{t\})^2\} = \mathfrak{M}\{(t - \theta)^2\}$. Under very general conditions it may be shown[1] that, for any fixed $\theta_1, \cdots, \theta_k$ and n, $\sigma^2\{t\}$ is always larger than or equal to a certain fixed number. In particular, if there is only one parameter we have for any *unbiased* estimate

$$\sigma^2\{t\} \geqq {\sigma_0}^2 = -\frac{1}{n\mathfrak{M}\left\{\frac{\partial^2 \ln \varphi}{\partial \theta^2}\right\}} > 0, \tag{2}$$

where φ in the continuous case is the probability density, $\varphi = \varphi(x; \theta)$; in the discontinuous case the probability, $\varphi = \varphi_i(\theta)$. If an (unbiased) estimate exists such that in (2) we have equality, this estimate is obviously the best possible; it is called an **efficient** estimate, t_{eff}. (From (2) and $\mathfrak{M}\{t_{\text{eff}}\} = \theta$ it follows that t_{eff} is also consistent.) For other unbiased estimates we have

$$0 \leqq \frac{{\sigma_0}^2}{\sigma^2\{t\}} = e\{t\} \leqq 1, \tag{3}$$

in which e is called the **efficiency** of t.

Sometimes it may be inconvenient to use an efficient estimate because it may involve cumbersome computations, but in that case the dispersion will be larger and the estimate obtained more inaccurate unless this is compensated for by increasing the number of observations.

Example. If in Kapteyn's distribution (10.3.2) we consider σ as known and μ as an unknown parameter to be estimated, (2) gives

$${\sigma_0}^2 = \frac{-1}{n\mathfrak{M}\left\{\frac{\partial^2}{\partial \mu^2}\left(-\ln(\sqrt{2\pi}\,\sigma) - \frac{(G(x) - \mu)^2}{2\sigma^2} + \ln G'(x)\right)\right\}} = \frac{-1}{n\mathfrak{M}\left\{\frac{-1}{\sigma^2}\right\}} = \frac{\sigma^2}{n}. \tag{4}$$

[1] With respect to the conditions for the validity of the theorems mentioned in this topic, as well as their proofs, see Cramér, *Mathematical Methods of Statistics*, especially Chapters 32–34.

Taking as an estimate of μ the average value (7.7.4), i.e., $m = \overline{G(x)}$, we see by comparing (7.7.7) with (4) that m is an unbiased, efficient estimate of μ. Other possible estimates of μ have therefore a larger dispersion than m, which fact may be checked directly in concrete cases.

***Exercise 1.** If in Kapteyn's distribution (10.3.2) we consider μ as known and $\theta = \sigma^2$ as an unknown parameter to be estimated, then show that (2) gives

$$\sigma^2\{t\} \geqq \sigma_0^2 = \frac{2\sigma^4}{n}. \tag{5}$$

Let

$$s_0^2 = \overline{(G(x) - \mu)^2} = \frac{(G(x_1) - \mu)^2 + \cdots + (G(x_n) - \mu)^2}{n}. \tag{6}$$

Then by means of Appendix 1, and by putting $y = G(x)$ as a new variable, show that

$$\mathfrak{M}\{s_0^2\} = \sigma^2, \qquad \sigma^2\{s_0^2\} = \frac{2\sigma^4}{n}. \tag{7}$$

Thus s_0^2 is an unbiased, efficient estimate of $\theta = \sigma^2$. Next let

$$s_1^2 = \overline{(G(x) - m)^2} = \frac{(G(x_1) - m)^2 + \cdots + (G(x_n) - m)^2}{n}, \tag{8}$$

where $m = \overline{G(x)}$. Putting $y = G(x)$, show that ns_1^2/σ^2 has the distribution (7.8.2) with $f = n - 1$. Next show from (7.8.3) that

$$\mathfrak{M}\{s_1^2\} = \frac{n-1}{n}\sigma^2, \qquad \sigma^2\{s_1^2\} = \frac{n-1}{n}\frac{2\sigma^4}{n}. \tag{9}$$

Thus $s^2 = n/(n-1)\, s_1^2$ (also introduced in (7.7.16)) is an unbiased estimate of σ^2. Show that s has the efficiency $(n-1)/n$.

***Exercise 2.** If in Kapteyn's distribution (10.3.2) we consider μ as known, and $\theta = \sigma$, and not σ^2, as an unknown parameter to be estimated, show that (2) gives

$$\sigma^2\{t\} \geqq \sigma_0^2 = \frac{\sigma^2}{2n}. \tag{10}$$

Putting $y = G(x)$ show that $\sqrt{n}\, s_0$, s_0 defined in (6), has the distribution (7.7.12) with $f = n$ (cf. Exercise 3, §7.8). By means of (7.7.14) and (7.7.15) show that

$$s_0' = \sqrt{\frac{n}{2}}\,\frac{\left(\frac{n-2}{2}\right)!}{\left(\frac{n-1}{2}\right)!}\, s_0 \sim \sqrt{\frac{n}{n-\frac{1}{2}}}\, s_0 \tag{11}$$

is an unbiased estimate of σ with dispersion

$$\sigma^2\{s_0'\} = \left(\frac{n}{2}\left(\frac{\left(\frac{n-2}{2}\right)!}{\left(\frac{n-1}{2}\right)!}\right)^2 - 1\right)\sigma^2 = \frac{\sigma^2}{2n} + O\left(\frac{1}{n^2}\right). \tag{12}$$

Thus the efficiency $e\{s_0'\} \xrightarrow[n\to\infty]{} 1$, and rather rapidly. Show that, for $n = 2$, (12) gives $e\{s_0'\} = \frac{\pi}{4(4-\pi)} = 0.915$ and, for $n = 3$, $e\{s_0'\} = \frac{4}{3(3\pi - 8)} = 0.936$. Show that s_1 defined in (8) has the distribution (7.7.12) with $f = n - 1$. From (7.7.14) and (7.7.15) then show that

$$s_1' = \sqrt{\frac{n}{2}}\,\frac{\left(\frac{n-3}{2}\right)!}{\left(\frac{n-2}{2}\right)!}\,s_1 \sim \sqrt{\frac{n}{n-\frac{3}{2}}}\,s_1 = \sqrt{\frac{n-1}{n-\frac{3}{2}}}\,s \tag{13}$$

is also an unbiased estimate of σ with dispersion

$$\sigma^2\{s_1'\} = \left(\frac{n}{2}\left(\frac{\left(\frac{n-3}{2}\right)!}{\left(\frac{n-2}{2}\right)!}\right)^2 - 1\right)\sigma^2 = \frac{\sigma^2}{2n} + O\left(\frac{1}{n^2}\right). \tag{14}$$

Thus the efficiency $e\{s_1'\} \xrightarrow[n\to\infty]{} 1$, but slower than $e\{s_0'\}$. Show that, for $n = 2$, (14) gives $e\{s_1'\} = \frac{1}{2(\pi - 2)} = 0.438$ and, for $n = 3$, $e\{s_1'\} = \frac{\pi}{6(4-\pi)} = 0.610$. Finally show that, for

$$s_2 = \sqrt{\frac{\pi}{2}}\,|G(x) - \mu| = \sqrt{\frac{\pi}{2}}\,\frac{|G(x_1) - \mu| + \cdots + |G(x_n) - \mu|}{n}, \tag{15}$$

we have

$$\mathfrak{M}\{s_2\} = \sigma, \qquad \sigma^2\{s_2\} = (\pi - 2)\frac{\sigma^2}{2n}. \tag{16}$$

Thus s_2 is also an unbiased estimate of σ, but with the efficiency $1/(\pi - 2) = 0.876$.

Exercise 3. We see from Exercises 1 and 2 that, if t is an unbiased, efficient estimate of θ, then $f(t)$ need *not* be an unbiased, efficient estimate of $f(\theta)$. However, show that if t is a consistent estimate of θ, i.e., $t \xrightarrow[n\to\infty]{\text{in } p} \theta$, then $f(t)$ is also a consistent estimate of $f(\theta)$, i.e., $f(t) \xrightarrow[n\to\infty]{\text{in } p} f(\theta)$.

Exercise 4. If in the binomial distribution (4.3.4) we consider ν as known and θ as an unknown parameter to be estimated, show that (2) gives

$$\sigma^2\{t\} \geqq \sigma_0^2 = \frac{\theta(1-\theta)}{n\nu}. \tag{17}$$

If

$$t = \frac{\bar{x}}{\nu} = \frac{x_1 + \cdots + x_n}{\nu n}, \tag{18}$$

then show that t is an unbiased, efficient estimate of θ.

Exercise 5. If in Poisson's distribution (4.3.5) μ is an unknown parameter to be estimated, then show that (2) gives

$$\sigma^2\{t\} \geqq \sigma_0^2 = \frac{\mu}{n}. \tag{19}$$

If

$$m = \tilde{x} = \frac{x_1 + \cdots + x_n}{n}, \tag{20}$$

then show that m is an unbiased, efficient estimate of θ.

***§ 10.10.** In practice two methods are used to obtain consistent estimates, although in special cases other estimates may also be convenient. In this topic we shall discuss the **method of maximum likelihood,** which was essentially developed by R. A. Fisher, although it was previously used by Gauss in a special case. In § 10.12 we shall discuss the **moment method,** essentially introduced by K. Pearson.

First we form the probability for the given sample, assuming that the n observations $x_1, \cdots, x_n$ are independent. In the discontinuous case this probability is given by

$$P(x_1, \cdots, x_n) = \varphi_{x_1}\varphi_{x_2} \cdots \varphi_{x_n}, \tag{1}$$

and in the continuous case we have

$$P(x_1, \cdots, x_n)\, dx_1 \cdots dx_n =$$

$$\varphi(x_1)\varphi(x_2) \cdots \varphi(x_n)\, dx_1 \cdots dx_n. \tag{2}$$

In both cases P will be a function of $x_1, \cdots, x_n$ as well as of the parameters $\theta_1, \cdots, \theta_k$. We now consider $x_1, \cdots, x_n$ as fixed and $\theta_1, \cdots, \theta_k$ as variables. The function

$$P(x_1, \cdots, x_n; \theta_1, \cdots, \theta_k) \geqq 0 \tag{3}$$

is then in both cases called the **likelihood function.** Now it is obvious that, if $\theta_1, \cdots, \theta_k$ had such values that P would be a very small number, then from **A,** § 9.2, we would not have expected to find the sample actually observed. Such a hypothesis regarding $\theta_1, \cdots, \theta_k$ has therefore to be discarded as unlikely. *Now the method of maximum likelihood simply asserts that the best possible estimates of* $\theta_1, \cdots, \theta_k$ *are those non-constant values which maximize the likelihood function.* In other words, the best estimates $t_1(x_1, \cdots, x_n), \cdots, t_k(x_1, \cdots, x_n)$ for $\theta_1, \cdots, \theta_k$ are obtained from the condition that P = maxi-

mum or, what is the same, that

$$\ln P = L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_k) = \text{maximum}. \tag{4}$$

Thus, in general, $t_1, \cdots, t_k$ are obtained as the solutions of the k **likelihood equations**[1]

$$\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial \theta_2} = \cdots = \frac{\partial L}{\partial \theta_k} = 0. \tag{5}$$

The solutions of the likelihood equations which are non-constants are called the **maximum likelihood estimates** and have under very general conditions the following properties (for the case of only one parameter, $k = 1$):

Theorem I. *If an efficient (and thus unbiased and consistent) estimate t_{eff} of θ exists, then the likelihood equation has a unique solution equal to t_{eff}.*

Theorem II. *The likelihood equation has a solution t_{lik} which is a consistent estimate of θ, i.e., $t_{\text{lik}} \underset{n\to\infty}{\overset{\text{in } p}{\longrightarrow}} \theta$. This solution is asymptotically normally distributed with the parameters $\mu = \theta$ and $\sigma^2 = \sigma_0^2$ given in* (10.9.2).

In theorem II the maximum likelihood estimate t_{lik} is said to be an **asymptotically efficient estimate** of θ. We stress that neither $\mathfrak{M}\{t_{\text{lik}}\}$ nor $\sigma^2\{t_{\text{lik}}\}$ need exist for any finite values of n. Furthermore, that if $\mathfrak{M}\{t_{\text{lik}}\}$ exists it need not be equal to θ; i.e., t_{lik} need not be unbiased. Sometimes it may be convenient to make it unbiased before using it, or perhaps to multiply it with another factor, g_n, such that $g_n \sim 1$ and $g_n \underset{n\to\infty}{\longrightarrow} 1$ (which will leave the *relative* dispersion unchanged).

Exercise 1. Show that (10.9.2) may also be written

$$\sigma_0^2 = \frac{-1}{\mathfrak{M}\left\{\dfrac{\partial^2 L}{\partial \theta^2}\right\}}, \tag{6}$$

where L is defined as in (4).

Example 1. For the binomial distribution (4.3.4) the likelihood function is

$$P(x_1, \cdots, x_n; \theta) = \binom{\nu}{x_1} \theta^{x_1} (1-\theta)^{\nu - x_1} \cdot \binom{\nu}{x_2} \theta^{x_2} (1-\theta)^{\nu - x_2} \cdots \binom{\nu}{x_n} \theta^{x_n} (1-\theta)^{\nu - x_n}. \tag{7}$$

[1] If φ has discontinuity points with respect to the parameters it may happen that one of these points gives the maximum value of L (cf. Problem 55).

We here consider ν as known, which is usually true, and the maximum likelihood estimate of θ is then from (5) given by

$$\frac{\partial L}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\ln\left(\prod_{i=1}^{n}\binom{\nu}{x_i}\right) + (x_1 + \cdots + x_n)\ln\theta + (n\nu - (x_1 + \cdots + x_n))\ln(1-\theta)\right] =$$

$$\frac{x_1 + \cdots + x_n}{\theta} - \frac{n\nu - (x_1 + \cdots + x_n)}{1-\theta} = 0,$$

i.e.,

$$\theta \approx t = \frac{x_1 + \cdots + x_n}{n\nu} = \frac{\bar{x}}{\nu}. \tag{8}$$

Check theorems I and II (cf. the result found in Exercise 4, § 10.9).

Example 2. For Poisson's distribution (4.3.5) the likelihood function is

$$P(x_1, \cdots, x_n; \mu) = \frac{e^{-\mu}\mu^{x_1}}{x_1!}\cdot\frac{e^{-\mu}\mu^{x_2}}{x_2!}\cdots\frac{e^{-\mu}\mu^{x_n}}{x_n!}. \tag{9}$$

From (5) we then obtain the maximum likelihood estimate of μ

$$\frac{\partial L}{\partial \mu} = \frac{\partial}{\partial \mu}\left[-n\mu + (x_1 + \cdots + x_n)\ln\mu - \ln\left(\prod_{i=1}^{n} x_i!\right)\right] =$$

$$-n + \frac{x_1 + \cdots + x_n}{\mu} = 0,$$

i.e.,

$$\mu \approx m = \frac{x_1 + \cdots + x_n}{n} = \bar{x}. \tag{10}$$

Check theorems I and II (cf. the result found in Exercise 5, § 10.9).

Example 3. Often the possible values 0, 1, 2, $\cdots$, will occur several times among the values x_1, x_2, $\cdots$ in the expressions (8) or (10). The calculation may therefore be simplified by counting the number of times, n_i, the result i occurs. Thus $\sum_i n_i = n$, and

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_i n_i i}{\sum_i n_i} \tag{11}$$

For example, inserting the values from the example, § 10.4, we get from (11)

$$\mu \approx m = \frac{10097}{2608} = 3.87.$$

Theorems I and II may be generalized to the case of more than one parameter. If we first consider the special case in which

$$\mathfrak{M}\left\{\frac{\partial^2 L}{\partial\theta_i\,\partial\theta_k}\right\} = 0 \qquad \text{for all} \qquad i \neq k, \tag{12}$$

then the generalization of (10.9.2) becomes: if $\theta_i \approx t_i$ then

$$\sigma^2\{t_i\} \geqq \sigma_{0i}^2 = \frac{-1}{n\mathfrak{M}\left\{\dfrac{\partial^2 \ln \varphi}{\partial\theta_i^2}\right\}} > 0, \qquad i = 1, 2, \cdots, k. \tag{13}$$

Exercise 2. Show that (13) may also be written

$$\sigma_{0i}^2 = \frac{-1}{\mathfrak{M}\left\{\dfrac{\partial^2 L}{\partial\theta_i^2}\right\}}, \tag{14}$$

where L is defined as in (4).

A set of estimates $t_1, \cdots, t_k$ for which the equality hold in (13) for all the t_i's is called a set of **joint efficient estimates.** Substituting in I and II "joint efficient" for "efficient," I and II also hold for more than one parameter, with σ_{0i}^2 given in (13), if (12) is fulfilled.

If (12) is not fulfilled we have to consider the moment matrix $\boldsymbol{M}_0^{(r)}$ of a set of joint efficient estimates, $t_1, \cdots, t_k$ (cf. Example 2, § 6.4), and in (13) σ_{0i}^2 is then to be replaced by the diagonal elements of $\boldsymbol{M}_0^{(r)}$, which is given by the natural generalization of (14):

$$\boldsymbol{M}_0^{(r)} = \mathfrak{M}\{t_r - \theta_r)(t_s - \theta_s)\} = -\left\{\mathfrak{M}\left\{\frac{\partial^2 L}{\partial\theta_r\,\partial\theta_s}\right\}\right\}^{-1}. \tag{15}$$

Exercise 3. Show that under the condition (12) eq. (15) reduces to (14).

Example 4. For Kapteyn's distribution (10.3.2) the likelihood function is

$$P(x_1, \cdots, x_n; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \cdot \exp\left[-\frac{1}{2\sigma^2}\left((G(x_1) - \mu)^2 + \cdots + (G(x_n) - \mu)^2\right)\right] \cdot G'(x_1) \cdots G'(x_n). \tag{16}$$

From (5) the maximum likelihood estimate of μ is given by

$$\frac{\partial L}{\partial \mu} = \frac{\partial}{\partial \mu}\left[-n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (G(x_i) - \mu)^2 + \ln \prod_{i=1}^{n} G'(x_i)\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n} (G(x_i) - \mu) = 0,$$

i.e.,

$$\mu \approx m = \frac{G(x_1) + \cdots + G(x_n)}{n} = \overline{G(x)}. \tag{17}$$

For the maximum likelihood estimate of σ we find

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} (G(x_i) - \mu)^2 = 0,$$

i.e.,

$$\sigma \approx s_1 = \left(\frac{1}{n}\sum_{i=1}^{n} (G(x_i) - m)^2\right)^{1/2} = \left(\overline{(G(x) - m)^2}\right)^{1/2}. \tag{18}$$

Finally, since $\mathfrak{M}\{G(x)\} = \mu$ and $\mathfrak{M}\{(G(x) - \mu)^2\} = \sigma^2$,

$$\mathfrak{M}\left\{\frac{\partial^2 L}{\partial \mu^2}\right\} = \mathfrak{M}\left\{-\frac{n}{\sigma^2}\right\} = -\frac{n}{\sigma^2},$$

$$\mathfrak{M}\left\{\frac{\partial^2 L}{\partial \sigma^2}\right\} = \mathfrak{M}\left\{\frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n} (G(x_i) - \mu)^2\right\} = -\frac{2n}{\sigma^2}, \tag{19}$$

and

$$\mathfrak{M}\left\{\frac{\partial^2 L}{\partial \sigma\, \partial \mu}\right\} = \mathfrak{M}\left\{-\frac{2}{\sigma^3}\sum_{i=1}^{n} (G(x_i) - \mu)\right\} = 0.$$

So from (15)

$$\sigma_{0\mu}{}^2 = \frac{\sigma^2}{n}, \qquad \sigma_{0\sigma}{}^2 = \frac{\sigma^2}{2n}. \tag{20}$$

Check theorems I and II (cf. the results of the example and exercises 1 and 2, § 10.9).

Exercise 4. If in Kapteyn's distribution we consider, besides μ, $\theta = \sigma^2$ and not σ itself as the parameter to be estimated, show that the maximum likelihood estimate of σ^2 is $s_1{}^2$ where s_1 given in (18) is the maximum likelihood estimate of σ.

Generalizing this result, show that, if $\theta \approx t_{\text{lik}}$, then $f(\theta) \approx f(t_{\text{lik}})$. We note that, if t is a consistent estimate of θ, then $f(t)$ is also a consistent estimate of $f(\theta)$, but that the properties unbiasness and efficiency may be destroyed by such a transformation (cf. Exercise 3, §10.9).

In the previous examples we have seen that, if a parameter was just the mean, μ, the best estimate m of μ was the average, $m = \bar{x}$. However, this is not always true.

Example 5. As a striking counter example we shall mention Cauchy's distribution (4.4.9). As discussed in Example 3, §8.1, $\bar{x}$ has here the same distribution as x, and so $m = \bar{x}$ is not even a consistent estimate of μ; as a matter of fact it does not contain any more information about μ than a single measurement does.

Exercise 5. For Laplace's distribution (4.4.10), with $\mathfrak{M}\{x\} = \mu, \sigma\{x\} = \sqrt{2}\,\alpha$, show that the best estimates of μ and α are $\mu \approx m = x_M$ and $\alpha \approx a = \dfrac{|x_1 - x_M| + \cdots + |x_n - x_M|}{n}$, where x_M denotes the sample median (cf. p. 55), i.e., if $x_1 \leqq x_2 \leqq \cdots \leqq x_n$, then for n odd $x_M = x_k$ for $k = (n+1)/2$, and for n even x_M may be any number between $x_{n/2}$ and $x_{(n/2)+1}$.

*§ **10.11.** For certain distributions there exist so-called **sufficient** estimates, i.e., estimates which for any value of n contain all the information about the parameter concerned, which it is at all possible to extract from the sample $x_1, \cdots, x_n$. Mathematically expressed such an estimate, t_s, has the property that, if in the likelihood function (10.10.3) we transform from the variables $x_1, \cdots, x_n$ to $t_s = t_s(x_1, \cdots, x_n)$ and $n-1$ other variables $u_1(x_1, \cdots, x_n), \cdots, u_{n-1}(x_1, \cdots, x_n)$, then in the case of only one parameter θ the parameter does *not* enter in the conditioned probability distribution of $u_1, \cdots, u_{n-1}$ under the condition that t_s has assumed the value t_s. In other words, the likelihood function may be written

$$P(x_1, \cdots, x_n; \theta) = g(t_s; \theta)h(u_1, \cdots, u_{n-1}|t_s), \tag{1}$$

where g gives the marginal distribution of t_s, which depends on θ, and h gives the conditioned distribution of $u_1, \cdots, u_{n-1}$, which does not depend on θ. Thus, if the value of t_s is known, the knowledge of the values of $u_1, \cdots, u_{n-1}$ cannot give us any further information about θ, which means that all the information about θ contained in the sample is also contained in t_s.

The generalization to the case of more than one parameter is obvious: a set of estimates $t_{s1}, \cdots, t_{sl}$ of $\theta_1, \cdots, \theta_l$ $(l \leqq k)$ is called a set of **joint sufficient** estimates of $\theta_1, \cdots, \theta_l$, if it is possible to transform from $x_1, \cdots, x_n$ to $t_{s1}(x_1, \cdots, x_n), \cdots, t_{sl}(x_1, \cdots,$

track belongs to either fragment, the distribution of x is a socalled *double Poisson distribution*

$$\varphi_i = \frac{1}{2}\frac{e^{-\mu'}\mu'^i}{i!} + \frac{1}{2}\frac{e^{-\mu''}\mu''^i}{i!}, \qquad i = 0, 1, 2, \cdots. \tag{4}$$

Here we have two parameters to estimate, μ' and μ'', and thus we have to calculate the first two moments of (4). From (5.1.6) and (5.3.11) we find

$$\mu_0 = 1, \quad \mu_1 = \tfrac{1}{2}\mu' + \tfrac{1}{2}\mu'', \quad \mu_2 = \mu_1 + \tfrac{1}{2}\mu'^2 + \tfrac{1}{2}\mu''^2. \tag{5}$$

The solution of these equations is readily found to be

$$\left.\begin{matrix}\mu' \\ \mu''\end{matrix}\right\} = \mu_1 \pm \sqrt{\mu_2 - \mu_1 - \mu_1^2}. \tag{6}$$

In Table 1 we give the distribution of $n = 327$ uranium fission fragments, n_i giving the number of tracks with i branches.[1] From

TABLE 1

i	n_i	$n_i i$	$n_i i^2$
0	28	0	0
1	47	47	47
2	81	162	324
3	67	201	603
4	53	212	848
5	24	120	600
6	13	78	468
7	8	56	392
8	3	24	192
9	2	18	162
10	1	10	100
	327	928	3736

the table we find the empirical moments

$$\begin{aligned}\mu_1 &\approx m_1 = \frac{1}{327}\sum_{i=0}^{10} n_i i = 2.8379 \\ \mu_2 &\approx m_2 = \frac{1}{327}\sum_{i=0}^{10} n_i i^2 = 11.425.\end{aligned} \tag{7}$$

[1] Bøggild, Broström, and Lauritzen, *Danske Vid. Selsk. Mat.-fys. Medd.*, Vol. XVIII, No. 4, 1940.

Equation (6) then gives

$$\mu' \approx 3.5684 \qquad \text{and} \qquad \mu'' \approx 2.1075 \tag{8}$$

Thus the two mean values are definitely different.

Exercise 3. Compute from (4) and (8) the theoretical distribution $\nu_i = n\varphi_i$, and show graphically that the agreement with the empirical distribution of Table 1 is satisfactory. Also compute the single Poisson distribution (4.3.5) for which from (7) $\mu \approx m_1 = 2.8379$, and show that this distribution agrees only poorly with the empirical distribution of Table 1.

*__Exercise 4.__ The distribution (4) is only a special case of the general double Poisson distribution

$$\varphi_i = \gamma_1 \frac{e^{-\mu'}\mu'^i}{i!} + \gamma_2 \frac{e^{-\mu''}\mu''^i}{i!}, \qquad \gamma_1 + \gamma_2 = 1, \tag{9}$$

having three independent parameters to estimate, μ', μ'', and γ_1.

Introducing the abbreviations

$$\beta_1 = \mu_1, \quad \beta_2 = \mu_2 - \mu_1, \quad \beta_3 = \mu_3 - 3\mu_2 + 2\mu_1, \tag{10}$$

show that μ'' is a root of the equation

$$(\beta_2 - \beta_1^2)x^5 - (\beta_3 - \beta_1^3)x^4 + 2(\beta_1\beta_3 - \beta_2^2)x^3 + 2\beta_3(\beta_2 - \beta_1^2)x^2 - (\beta_3^2 - \beta_2^3)x + \beta_3(\beta_1\beta_3 - \beta_2^2) = 0. \tag{11}$$

Next show that μ', γ_1, and γ_2 are given by

$$\mu' = \frac{\beta_3 - \beta_1\mu''^2}{\beta_2 - \mu''^2}, \quad \gamma_1 = \frac{\mu'' - \beta_1}{\mu'' - \mu'}, \quad \gamma_2 = \frac{\beta_1 - \mu'}{\mu'' - \mu'}. \tag{12}$$

*__Exercise 5.__ Denoting by b_1, b_2, b_3 the corresponding empirical values of β_1, β_2, β_3 find from Table 1 the values

$$m_3 = \frac{1}{327}\sum_{i=0}^{10} n_i i^3 = 56.398$$
$$b_1 = 2.8379, \quad b_2 = 8.5872, \quad b_3 = 27.798, \tag{13}$$

and next deduce the values

$$\gamma_1 \approx 0.27413, \qquad \gamma_2 \approx 0.72587,$$
$$\mu' \approx 4.0276, \qquad \mu'' \approx 2.3886. \tag{14}$$

Finally compute from (9) and (14) the theoretical distribution $\nu_i = n\varphi_i$, and show graphically that it fits the empirical distribution of Table 1 just as well as the special double Poisson distribution for which $\gamma_1 = \gamma_2 = \frac{1}{2}$. Thus empirically we cannot in this case distinguish between the general and the special form. However, for theoretical reasons the latter is to be preferred.

11.

APPLICATION OF THE THEORY OF PROBABILITY TO THE THEORY OF ERRORS

§11.1. By the theory of errors—a somewhat unfortunate but now generally adopted name—is understood that special branch of statistics which deals with the numerical determination of physical quantities. However, since the normal distribution is always used here, the methods of this chapter may be applied whenever the normal distribution, or Kapteyn's generalization of it, may be regarded as giving a satisfactory description of the statistical phenomena considered.

The four questions considered in the theory of errors are the following: I. What is to be understood by the "true" value of a physical quantity? II. In practice, how can the true values be estimated from measurements? III. What certainty can we attribute to the estimates? IV. How can we compare estimates obtained from different sets of measurements?

§11.2. Before answering question I, we must first of all realize that a *physical* quantity, as defined by a certain measuring prescript, is, in contrast to a mathematical quantity, never so exactly defined that any one definite number rather than all other numbers can be regarded as giving the "true" value of the quantity. Because of the limited perfection of our senses and of all our measuring instruments there is always a certain limitation as to how small a quantity we can observe. As a result of this fact we can state only a finite, although larger or smaller, number of digits. Or, in other words, *every empirically formed number is an integer when expressed in units of the smallest unit, ϵ, that the measuring apparatus can measure.* Thus any real number within an interval of length ϵ may be called the "true" value of the quantity. Furthermore, this uncertainty in definition is, as a rule, small as compared to the uncertainties arising from the measuring errors which we have discussed in §1.2. Nevertheless, since an indefinite improvement of the accuracy of our measuring technique seems possible in principle, it is for a mathematical description a natural idea to abstract from these uncertainties and to idealize our observations by selecting one definite number as *the* true value of the quantity considered. It

is just these true values, and not the directly observed values, which enter into our model of the reality (§1.1), viz., the laws of classical physics. In quantum theory, this idealization has turned out to be too coarse for atomic phenomena. Instead of such a causal description in which all quantities have definite values, we must in our model also use a statistical description by means of distribution functions differing from the causal distribution (cf. §4.17). Comparing our discussions of the concepts "true value" and "probability," we see that "probability" is not different from the "true value" of any other physical quantities. Probabilities are simply the true values of special quantities, viz., relative frequencies.

§11.3. We shall now neglect the uncertainties inherent in the definitions proper and discuss only those arising from the measurements. As discussed in §1.2 the imperfections of our measuring instruments as well as disturbing factors always present result in the fact that every measured result is a *random variable*, because repeated measurements of one and the same quantity, of the constancy of which we feel convinced, will yield different values. We express this fact by saying that the measurements are encumbered with **errors** causing the values measured to deviate more or less from the true values. This will, of course, be the case only if the measuring instruments are not too coarse. If, e.g., the distance between two fine marks on a steel rod is 2.0078 m and we measure it with a rule which has only divisions at every meter, each measurement will give the result "2 m." Therefore it is always assumed that the measuring instrument is chosen such that the smallest unit it can measure is small compared with the quantity measured.

Errors are divided into three groups: (I) *Coarse or gross errors.* (II) *Systematic errors.* (III) *Statistical or random errors.*

Coarse errors are errors in reading the instruments, in computations, caused by wrong treatment of the instruments, or simply by lack of care on the observer's part. Of course, such errors should be avoided. Observations encumbered with coarse errors are usually immediately conspicuous because they have values quite different from the other observations. (In ordnance one speaks of stray shots.) In the following we shall assume that observations encumbered with coarse errors have already been discarded (however, cf. §11.17).

Systematic errors are errors due to one or to a few definite causes acting according to a definite law and, as a rule, in one definite direction. If a measurement is repeated under constant conditions, the same systematic errors will occur. Consequently, in contrast to the coarse errors, systematic errors will not show up in any disagreement

among the different results, but only displace them by a constant amount. However, if the laws governing the systematic errors are known, these errors can be calculated and treated as *corrections* to the values measured. Most systematic errors are caused by the instruments. If, e.g., a length is measured with a rule a little shorter than its divisions state, each measurement will give too large a result. However, this error will not be noticed unless the measurement is repeated with a second rule. Such comparisons between the results obtained by different measuring instruments is the most efficient method for detecting systematic errors. In the following we shall assume that our measurements have been corrected for known systematic errors.

Random errors are all the other errors which do not show any regularities or the regularities of which we do not know. Sometimes the word errors is applied only to the systematic errors, the word uncertainties to the random errors. Most random errors arise from the interpolation necessary in reading scales, from the adjustment of the measuring instruments, from the manufacturing of the scales and standards, and furthermore from all the disturbing factors as discussed in § 1.2. In general, it is a characteristic feature of random errors, in contrast to the systematic errors, that positive and negative values are equally probable. However, errors having skew distributions may be found; these are the so-called *one-sided* errors. By closer investigation one-sided errors often turn out to be systematic errors. As an example of a one-sided error we may mention the curvature in the axes of optical instruments.

The distinction between the various groups of errors is, however, not sharp. By closer investigation some of the random errors may show regularities. Thus an error we have previously classified as random may later turn out to be systematic. Casually there may also appear a particularly large error which may be mistaken for a coarse error. Thus in practice it is not always easy to judge whether or not a measurement which deviates conspicuously from the other measurements should be rejected (cf. § 11.17).

§ 11.4. To a given physical quantity x and to a given measuring method we now associate a certain distribution function $\Phi(x)$ (cf. footnote 2, p. 137). We stress that $\Phi(x)$ *also depends on the measuring method*. In fact, a physical quantity is not defined before a method of measuring it has been stated. The four questions in § 11.1 cannot be answered independent of $\Phi(x)$. However, experience has shown that most physical measurements—corrected for coarse and systematic errors—are approximately normally distributed, a fact which may be

explained theoretically (cf. Example 1, §10.3). We shall therefore base the theory of errors on the normal distribution

$$d\Phi = \varphi(x)\,dx = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx, \tag{1}$$

although strictly speaking every physical quantity is a *discontinuous* variable. However, if the statistical fluctuations are large as compared with the smallest measurable unit, the distribution will have a large number of very small steps and thus may be approximated by a *continuous* distribution (cf. p. 32).

Since, first, the parameter μ is equal to the mean value which gives the order of magnitude of x (cf. IV, §9.3) and, second, the normal distribution is symmetic about $x = \mu$, it is natural to denote μ by definition as the "true" value of the physical quantity x. We stress that from this definition the true value depends also on the method of measurement applied. Thus it is an important task to decide whether or not two different values for the true value obtained by two different methods of measuring are in agreement within the statistical fluctuations (cf. §11.15). If they disagree significantly this fact shows either that we have not measured the same quantity or that certain systematic errors have been overlooked. We then obtain the true value when these systematic errors have been localized and taken into account.

§11.5. We shall now discuss the next question: how, from measurements, to estimate the numerical value of μ. Physical measurements fall into two groups: direct and indirect measurements. In a **direct** measurement such as the measurement of a length, a mass, or a current with an ammeter, the numerical result is obtained directly from the observations. In an **indirect** measurement such as the measurement of specific resistance, $\rho = \pi R d^2/4l$, of a wire with total resistance R, length l, and diameter d, the result has to be computed from the observations, the quantity to be measured being a function of certain other quantities which are measured directly. We shall first consider direct measurements exclusively, discussing indirect measurements in §11.22. To determine the true value, μ, of the quantity x considered, it is, because of the statistical fluctuations, insufficient to perform only one measurement of x. Only in those cases where it may be expected in advance, e.g., from earlier measurements, that the statistical fluctuations will be small is a single measurement, x_1, sufficient. According to the method of maximum likelihood, x_1 will then be the best estimate of μ. In general, the measurement must be repeated a certain number of times, giving the results $x_1, x_2, \cdots, x_n$. As shown in Example 4

§ 10.10, putting $G(x) = x$, in which case Kapteyn's distribution (10.3.2) reduces to the normal (11.4.1), the best estimate of the true value μ is

$$\mu \approx m = \bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{[x]}{n}. \tag{1}$$

As shown in the example, § 10.9, m is for all values of n an unbiased, efficient estimate of μ. In (1)

$$[x] = \sum_{i=1}^{n} x_i \tag{2}$$

is a symbol introduced by Gauss and still much used in the theory of errors. It is also useful in the theory of errors to introduce the "true errors,"[1] or the residuals,

$$\epsilon_i = x_i - \mu, \qquad i = 1, 2, \cdots, n \tag{3}$$

and the "best errors"

$$v_i = x_i - \bar{x}, \qquad i = 1, 2, \cdots, n. \tag{4}$$

Exercise 1. Show that

$$[v] = \sum_{i=1}^{n} v_i = 0. \tag{5}$$

Next show that

$$\sum_{i=1}^{n} (x_i - x')^2 = [v'v'] = [vv] + n(\bar{x} - x')^2, \tag{6}$$

where $v_i' = x_i - x'$ and x' is an arbitrary number.

Since μ appears in the likelihood function only in the expression $\sum_{i=1}^{n} (x_i - \mu)^2 = [\epsilon\epsilon]$ (cf. (10.10.16)), we see that for a normal distribution the best estimate of μ is obtained by the condition that the best errors satisfy

$$[vv] = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \text{minimum}. \tag{7}$$

[1] The *error* is the quantity that must be added to the "true value" in order to obtain the measured value; the error with sign reversed is called the *correction*:

$$\text{true value} + \text{error} = \text{measured value}$$

$$\text{measured value} + \text{correction} = \text{true value}.$$

This famous principle is called the **method of least squares.**[1] From (6) we see that $x' = \bar{x}$ really corresponds to the smallest possible value of $[v'v']$.

Exercise 2. For the best errors (4) show that

$$\mathfrak{M}\{v_i\} = 0, \qquad i = 1, 2, \cdots, n. \tag{8}$$

Furthermore show that

$$v_i = \sum_{j=1}^{n} \left(\delta_{ij} - \frac{1}{n}\right) x_j, \qquad \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{9}$$

and thus

$$\sigma^2\{v_i\} = \left(1 - \frac{1}{n}\right)^2 \sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = \frac{n-1}{n}\sigma^2, \qquad i = 1, 2, \cdots, n. \tag{10}$$

Finally show that

$$\mathfrak{M}\{[vv]\} = (n-1)\sigma^2 \tag{11}$$

in agreement with (7.7.13) since $q^2 = [vv]$.

From Exercise 2 it follows that each v_i, being a sum of normally distributed independent variables, is normally distributed with the parameters 0 and $\sqrt{\frac{n-1}{n}}\sigma$. However, because of the constraint (5) the v_i's are not independent.

***Exercise 3.** Show that

$$\rho\{v_i, v_j\} = \frac{-1}{n-1}, \qquad i \neq j \tag{12}$$

but

$$\rho\{v_i, \bar{x} - \mu\} = 0. \tag{13}$$

§11.6. As shown in Example 4, §10.10, the best estimate of the parameter σ is given by

$$\sigma \approx s_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{[vv]}{n}}. \tag{1}$$

However, as discussed in Exercise 2, §10.9, s_1 is only an asymptotically unbiased and efficient estimate of σ, and of course there exist many other estimates which are also asymptotically unbiased and efficient, e.g., all estimates of the form $g_n s_1$, where $g_n \sim 1$ and $g_n \xrightarrow[n\to\infty]{} 1$, which have the same *relative* dispersion as s_1. In statistical practice it is

[1] French, **méthode des moindres carrés;** German, **Methode der kleinsten Quadrate.**

customary to use s defined in (7.7.16), i.e., to choose $g_n = \sqrt{\frac{n}{n-1}}$:

$$\sigma \approx s = \sqrt{\frac{n}{n-1}}\, s_1 = \sqrt{\frac{[vv]}{n-1}}, \tag{2}$$

where $n - 1 = f$ is called the **degree of freedom.** Whether to use s_1 or s is a question of personal choice. However, as discussed in § 11.10, s has certain advantages over s_1, as a consequence of which s is, as a rule, preferred.

§ 11.7. The two estimates, $\bar{x}$ and s, for the parameters μ and σ are themselves random variables, being subject to statistical fluctuations, i.e., giving other values if the n measurements from which they have been calculated are repeated. In order to estimate their accuracy we next determine the dispersions of $\bar{x}$ and s. As discussed in § 7.5, p. 88, $\bar{x}$ is for all values of n normally distributed with mean value μ and dispersion $\sigma/\sqrt{n}$, i.e.,

$$\sigma\{\bar{x}\} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \sqrt{\frac{[vv]}{n(n-1)}}; \tag{1}$$

furthermore, we have seen that $\sqrt{n-1}\, s = q$ has the q-distribution (7.7.12), which for large values of n is approximately normal. From (7.7.18) and (7.7.19) s is approximately normally distributed with mean value σ and dispersion $\sigma/\sqrt{2(n-1)}$, so that

$$\sigma\{s\} \sim \frac{\sigma}{\sqrt{2(n-1)}} \approx \frac{s}{\sqrt{2(n-1)}} = \frac{\sqrt{[vv]}}{\sqrt{2}\,(n-1)}. \tag{2}$$

Example 1. We note that (2) shows that, if there were no other reasons to prefer s over s_1, they would be equally good, because for n-values that are not too small their difference is small as compared with their dispersion:

$$s - s_1 = \sqrt{[vv]}\left(\frac{1}{\sqrt{n-1}} - \frac{1}{\sqrt{n}}\right) =$$

$$s\left(1 - \left(1 - \frac{1}{n}\right)^{1/2}\right) = s\left(\frac{1}{2n} + \cdots\right),$$

which is just small as compared with $\sigma\{s\} \approx s\left(\frac{1}{\sqrt{2n}} + \cdots\right)$.

Another argument often presented for preferring s over s_1 is that s^2 is an unbiased estimate of σ^2, whereas s_1^2 is not (cf. Exercise 1, § 10.9).

However, first, it is more natural to consider σ and not σ^2 as the parameter to be estimated as long as we do not introduce another letter for σ^2; but in this case neither s nor s_1 is unbiased (cf. Exercise 2, §10.9). Second, both s and s_1 converge in probability towards σ, which means that for large n the probability mass of both estimates, having the same *relative* dispersions, is concentrated about σ; but in this case their difference is small compared to their dispersions, as shown above. On the other hand, $\sigma\{s\}$ tends so slowly towards 0 that the actually observed value need not lie very close to σ unless n is extremely large. The only reason for preferring s over s_1, if at all, is that given in §11.10.

The final result of the n measurements $x_1, \cdots, x_n$ is thus given as:
Estimate of the true value:

$$\mu \approx \bar{x} = \frac{[x]}{n} \qquad \textit{with dispersion:} \quad \sigma\{\bar{x}\} \approx \frac{s}{\sqrt{n}}. \tag{3}$$

Estimate of the dispersion:

$$\sigma \approx s = \sqrt{\frac{[vv]}{n-1}} \qquad \textit{with dispersion:} \quad \sigma\{s\} \approx \frac{s}{\sqrt{2(n-1)}}. \tag{4}$$

Often the dispersion of a statistic is called its **standard error, mean-square error,** or simply **mean error.**[1] The dispersion, σ, itself is, especially in older literature, called the *standard* or *mean error of a single measurement*, which is a rather unfortunate notation.

We stress that, to judge the values of $\bar{x}$ and s, their *dispersions should always be given*, or in any case the number, n, of measurements from which the values have been obtained.

Example 2. Sometimes the result is given in the form

$$\mu = \bar{x} \pm \frac{s}{\sqrt{n}}, \qquad \text{or} \qquad \mu = \bar{x} \pm 100\,\frac{s}{\bar{x}\sqrt{n}}\,\%. \tag{5}$$

However, first, we have not $\mu = \bar{x}$, but $\mu \approx \bar{x}$. Second, μ may very well have values outside this interval and still be compatible with the values $\bar{x}$ and s observed (cf. §11.9). Equation (5) is therefore very misleading and should be avoided, (3) being preferable.

As mentioned in §7.4 another parameter is sometimes used, especially in older literature, instead of σ, viz., $\rho = 0.67449\sigma$, which is the 50% tolerance limit. The result is then written

[1] French, **écart type;** German, **mittlere quadratische Fehler.**

$$\mu = \bar{x} \pm 0.674 \frac{s}{\sqrt{n}} = \bar{x} \pm \frac{r}{\sqrt{n}}, \qquad r = 0.674s. \tag{6}$$

Since a measurement is just as likely to fall inside as outside the 50% tolerance limit a statement like (6) is certain to lead to an underestimation of the accuracy of $\bar{x}$, and it should therefore never be used.

Example 3. Instead of s as given in (1) other estimates of σ are also used, e.g., s_2 given in (10.9.15). For a rough estimate, as is often sufficient in engineering, it is very convenient to find the largest value, $x_{\max}$, and the smallest, $x_{\min}$, among the n observations and then to put $\mu \approx m_1 = \frac{1}{2}(x_{\max} + x_{\min})$ and $\sigma \approx s_3 = \frac{1}{\alpha_n}(x_{\max} - x_{\min})$, in which α_n is determined so that $\mathfrak{M}\{s_3\} = \sigma$. ($w = x_{\max} - x_{\min}$ is called the **range**.) It may be shown that m_1 and s_3 are consistent, but very inefficient estimates of μ and σ. The function α_n has been tabulated.[1] For example, we find $\alpha_{10} = 3.08$, $\alpha_{30} = 4.09$, $\alpha_{100} = 5.02$, and $\alpha_{500} = 6.07$.

***Exercise.** An estimate sometimes used, e.g., in ordnance, is

$$s_4 = \frac{\sqrt{\pi}}{2} \frac{d_1 + \cdots + d_{n-1}}{n(n-1)}, \qquad d_i = |x_i - x_{i+1}|,$$

where the d_i's are called the **successive differences.** Show that s_4 is an unbiased estimate of σ. By a somewhat lengthy calculation it may be shown that s_4 is a consistent estimate with the efficiency $e\{s_4\} = 0.61 \frac{n-1}{n}$.

§11.8. If we know μ and σ, then the **tolerance limits,** i.e., the limits (assumed, as a rule, to lie symmetric about μ) *within* which x lies with a given probability $1 - P$, may be found for all values of P from Table II (cf. §7.4). Because of the statistical fluctuations of both $\bar{x}$ and s these values may not be substituted immediately for μ and σ in the expressions for the tolerance limits, although this is often done. However, in the t-distribution (7.9.3) we have a distribution which does not contain the parameters μ and σ themselves and from which we can estimate the tolerance limits of a new measurement, x_{n+1}, on the basis of $\bar{x}$ and s calculated from a sample $x_1, \cdots, x_n$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{1}$$

[1] See the discussion and literature in Cramér, *Mathematical Methods of Statistics*, p. 374. A table of α_n for $n = 1 - 1000$ is given in L. H. C. Tippett, *Biometrika*, **17**, 364, 1925.

We then form the new random variable

$$t = \frac{\bar{x} - x_{n+1}}{\sqrt{1 + \frac{1}{n}}\, s}, \tag{2}$$

which as shown in Exercise 2, §7.9, has the t-distribution (7.9.3) with $f = n - 1$. If $t(P, f)$ denotes the solution of the equation $P(|t| \geqq t) = P$, which is tabulated in Table III, then

$$P\left(-t(P, f) \leqq \frac{\bar{x} - x_{n+1}}{\sqrt{1 + \frac{1}{n}}\, s} \leqq t(P, f)\right) = 1 - P. \tag{3}$$

But this may also be written, solving the inequalities for x_{n+1},

$$P\left(\bar{x} - \sqrt{1 + \frac{1}{n}}\, s t(P, f) \leqq x_{n+1} \leqq \bar{x} + \sqrt{1 + \frac{1}{n}}\, s t(P, f)\right) = 1 - P. \tag{4}$$

Thus corresponding to a given value of P the tolerance limits of a new measurement, x_{n+1}, yet to be performed, are estimated from the given sample to be

$$\bar{x} - \sqrt{1 + \frac{1}{n}}\, s t(P, f), \qquad \bar{x} + \sqrt{1 + \frac{1}{n}}\, s t(P, f). \tag{5}$$

Since the t-distribution tends to the normalized normal distribution for $n \to \infty$, the factor $\sqrt{1 + \frac{1}{n}}\, t(P, f)$ will for $n \to \infty$ tend to the corresponding factor as obtained from Table II. In practice it is customary to choose $P = 5\%$.

Example. For the 10 measurements given in the example, §11.11, $\bar{x} = 4.0759$ and $s = 0.0039$. Since $f = 10 - 1 = 9$ and $t(5\%, 9) = 2.262$ from Table III, we have $\sqrt{1 + 1/10}\, t(5\%, 9) = 2.371$, which is considerably larger than $t(5\%, \infty) = 1.96$. Thus we find the tolerance limits for a new measurement to be $4.0759 \pm 0.0039 \cdot 2.371 = 4.0759 \pm 0.0092$.

§11.9. In (11.7.3) we have given the best estimate of the true value μ, $\bar{x}$, and its dispersion, $s/\sqrt{n}$. Now μ is a definite, although unknown, *constant*, i.e., a parameter and not a random variable. Thus we cannot speak of the probability of μ lying inside certain limits, say $\bar{x} \pm s/\sqrt{n}$. However, we may ask for the values of μ which are compatible with the observed values. To that purpose we may again use the t-dis-

tribution (7.9.3), forming the random variable

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}, \tag{1}$$

which as shown in Exercise 2, § 7.9, has the distribution (7.9.3) with $f = n - 1$. If μ is varied, so is the value of t, and it is now reasonable to consider a definite value of μ compatible with the sample observed, $x_1, \cdots, x_n$, if the corresponding t-value lies within reasonable limits, for which it is customary to take the 5% limits. Thus, if $t(P, f)$ has the same meaning as in § 11.8, μ is considered compatible with the sample if it satisfies the equation

$$P\left(-t(P, f) \leqq \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \leqq t(P, f)\right) = 1 - P. \tag{2}$$

Solving the inequalities for μ this may also be written

$$P\left(\bar{x} - \frac{s}{\sqrt{n}}\, t(P, f) \leqq \mu \leqq \bar{x} + \frac{s}{\sqrt{n}}\, t(P, f)\right) = 1 - P. \tag{3}$$

These limits of μ are called **confidence** or **fiducial limits.** We stress that the concept of confidence limits must not be confused with that of tolerance limits discussed in the preceding topic. The latter are probability limits for a random variable; the former, limits of compatibility with observations for a parameter.

Example. For the 10 measurements given in the example, § 11.11, $\bar{x} = 4.0759$ and $s/\sqrt{10} = 0.0012$. Since $f = 10 - 1 = 9$ and from Table III $t(5\%, 9) = 2.262$, we get the confidence limits of the true value $4.0759 \pm 0.0027 = (4.0732, 4.0786)$.

§ 11.10. Unfortunately it is not always practicable, or possible, to perform the great number of measurements necessary to get from (11.7.4) a reliable estimate of σ. However, we often have many short series of measurements belonging to different values of μ but taken under the same conditions and, therefore, having the same accuracy, i.e., the same value of σ. Since the distribution of $q = \sqrt{[vv]}$, (7.7.12), depends only on σ and $f(= n - 1)$, but not on μ, it may be convenient to treat q as the primary variable, but not the directly measured x_i's from which q has been calculated. In the latter case the n measurements, $x_1, \cdots, x_n$, represent n observations of the random variable $\boldsymbol{x}$, but in the former only *one* measurement of the random variable $\boldsymbol{q}$.

Thus we shall expect to find different estimates of the same parameter σ in the two cases.

Let us assume that we have m series of measurements of x, consisting of $n_1, \cdots, n_m$ measurements, respectively, i.e., with the degrees of freedom $f_1 = n_1 - 1, \cdots, f_m = n_m - 1$. Let $q = \sqrt{[vv]}$ from the m series be $q_1 = \sqrt{[vv]_{n_1}}, \cdots, q_m = \sqrt{[vv]_{n_m}}$, i..e, let $s_1 = q_1/\sqrt{f_1}$, $\cdots$, $s_m = q_m/\sqrt{f_m}$ be the m estimates of σ obtained from each series. Assuming, of course, the m series to be independent, the likelihood function (10.10.3) becomes, from (7.7.12),

$$P(q_1, \cdots, q_m; \sigma, f_1, \cdots, f_m) =$$

$$\frac{\text{const}}{\sigma^m}\left(\frac{q_1}{\sigma}\right)^{f_1-1} \cdots \left(\frac{q_m}{\sigma}\right)^{f_m-1} \exp\left[-\frac{1}{2\sigma^2}(q_1^2 + \cdots + q_m^2)\right]. \quad (1)$$

Thus the maximum likelihood estimate of σ is the solution of

$$\frac{\partial L}{\partial \sigma} = \frac{\partial}{\partial \sigma} \ln P = -\frac{f}{\sigma} + \frac{q^2}{\sigma^3} = 0 \quad (2)$$

$$f = f_1 + \cdots + f_m = n - m, \qquad q^2 = q_1^2 + \cdots + q_m^2,$$

i.e.,

$$\sigma \approx s = \frac{q}{\sqrt{f}} = \sqrt{\frac{[vv]_n}{n-m}} = \sqrt{\frac{(n_1-1)s_1^2 + \cdots + (n_m-1)s_m^2}{(n_1-1) + \cdots + (n_m-1)}}. \quad (3)$$

From (10.10.6) the asymptotic dispersion of s, which is asymptotically normally distributed with mean value σ, is given by

$$\sigma_0^2\{s\} = \frac{-1}{\mathfrak{M}\left\{\frac{\partial^2 L}{\partial \sigma^2}\right\}} = \frac{-1}{\mathfrak{M}\left\{\frac{f}{\sigma^2} - 3\frac{q^2}{\sigma^4}\right\}} = \frac{\sigma^2}{2f} \quad (4)$$

(check). Thus

$$\sigma\{s\} \sim \frac{\sigma}{\sqrt{2f}} \approx \frac{s}{\sqrt{2(n-m)}}. \quad (5)$$

For $m = 1$ these formulae reduce to (11.7.4). We see that, although for a single series, $m = 1$, it is unessential whether we divide by n or f, it is essential when we add several series for estimating σ, since it is the degrees of freedom which are added, not the number of measurements. *This is the only reason why it may be more convenient to prefer s over s_1.*

This method of estimating σ is very useful in practice and should, therefore, be used much more than it is in contemporary literature.

Thus we often repeat a measurement in order to check for coarse errors, and in such cases one double measurement is quite useless for estimating σ. But if we have a number of double measurements with the same accuracy, i.e., the same σ (but possibly different values of μ), they can nevertheless be used for an accurate estimate of σ.

§11.11. The numerical computation of $\bar{x}$ and s may be simplified by first guessing a preliminary value of $\bar{x}$, x^0. We then have

$$\bar{x} = \frac{[x]}{n} = \frac{[x^0 + v^0]}{n} = x^0 + \frac{[v^0]}{n}, \tag{1}$$

where

$$v_i{}^0 = x_i - x^0. \tag{2}$$

To check this calculation we may compute the best errors, $v_i = x_i - \bar{x}$, and check whether they satisfy (11.5.5), viz., $[v] = 0$. Next, putting $x' = x^0$ in (11.5.6) we have

$$[vv] = [v^0 v^0] - n(\bar{x} - x^0)^2 = [v^0 v^0] - \frac{[v^0]^2}{n}. \tag{3}$$

As a check we may again compute the best errors and see whether they satisfy

$$[vv] = [x^2] - \frac{[x]^2}{n}, \tag{4}$$

which is obtained from (11.5.6) by putting $x' = 0$.

Example. By 10 readings of a micrometer the results $x_1, \cdots, x_{10}$ given in the table are obtained. Using (1) and (3) the computations then run as follows:

i	x_i	$v_i{}^0$	$v_i{}^{02}$
1	4.078	3×10^{-3}	9×10^{-6}
2	4.080	5	25
3	4.071	−4	16
4	4.076	1	1
5	4.081	6	36
6	4.077	2	4
7	4.075	0	0
8	4.073	−2	4
9	4.079	4	16
10	4.069	−6	36
		9×10^{-3}	147×10^{-6}

$$x^0 = 4.075$$

$$[v^0] = 9 \times 10^{-3}$$

$$[v^0 v^0] = 147 \times 10^{-6}$$

$$[vv] = (147 - {}^{81}\!/_{10}) \times 10^{-6} = 138.9 \times 10^{-6}$$

$$\bar{x} = 4.075 + {}^{9}\!/_{10} \times 10^{-3} = 4.0759$$

$$s = \sqrt{\frac{138.9}{9}} \times 10^{-3} = 0.00393$$

$$\frac{s}{\sqrt{10}} = 0.00124$$

$$\frac{s}{\sqrt{18}} = 0.000926$$

Result:

$\mu \approx \bar{x} = 4.0759$ with standard error 0.0012

$\sigma \approx s = 0.0039$ with standard error 0.0009.

We see from this example that in most of the computations only small integers occur so that most of them may be performed mentally.

§ 11.12. If the observations have already been grouped and if n is so large that the Δt_i's are small, we may, to a good approximation, put all the values in a class equal to the value of the middle point of the class, t_i. Thus we obtain (cf. § 10.5)

$$\bar{x} \sim \frac{n_1 t_1 + \cdots + n_m t_m}{n_1 + \cdots + n_m} = \frac{[nt]}{[n]} \tag{1}$$

and

$$[vv] \sim n_1(t_1 - \bar{x})^2 + \cdots + n_m(t_m - \bar{x})^2 = [n(t - \bar{x})^2]. \tag{2}$$

If again we first make a preliminary guess as to $\bar{x}$, x^0 equal to one of the t_i's, t^0, we obtain again very simple computations, especially if all the class intervals are of equal length. In fact, putting $t_i - t^0 = r_i \Delta t$, we have

$$\bar{x} = t^0 + \Delta t \frac{[nr]}{n}$$

$$[vv] = (\Delta t)^2 \left([nr^2] - \frac{[nr]^2}{n} \right) \tag{3}$$

(check).

Example. For the 96 shots in the example, § 10.5, the computations run as follows, $\Delta t = 10$,

Azimuth Deviation

t_i	n_i	r_i	r_i^2	$n_i r_i$	$n_i r_i^2$
−30	0	−3	9	0	0
−20	2	−2	4	−4	8
−10	9	−1	1	−9	9
0	28	0	0	0	0
10	30	1	1	30	30
20	21	2	4	42	84
30	5	3	9	15	45
40	1	4	16	4	16
50	0	5	25	0	0
	96			78	192

$t^0 = 0$

$[nr] = 78$

$[nr^2] = 192$

$$[vv] = 100\left(192 - \frac{78^2}{96}\right) = 12{,}863$$

$$\bar{x} = 0 + \frac{10 \cdot 78}{96} = 8.125 (8.278)$$

$$s = \sqrt{\frac{12863}{95}} = 11.636 (11.393)$$

$$\frac{s}{\sqrt{96}} = 1.188$$

$$\frac{s}{\sqrt{190}} = 0.844$$

Height Deviation

t_i	n_i	r_i	r_i^2	$n_i r_i$	$n_i r_i^2$
−60	0	−6	36	0	0
−50	3	−5	25	−15	75
−40	5	−4	16	−20	80
−30	13	−3	9	−39	117
−20	18	−2	4	−36	72
−10	21	−1	1	−21	21
0	21	0	0	0	0
10	10	1	1	10	10
20	5	2	4	10	20
30	0	3	9	0	0
	96			−111	395

$t^0 = 0$

$[nr] = -111$

$[nr^2] = 395$

$$[vv] = 100\left(395 - \frac{111^2}{96}\right) = 26{,}666$$

$$\bar{x} = 0 - \frac{10 \cdot 111}{96} = -11.562 (-11.969)$$

$$s = \sqrt{\frac{26666}{95}} = 16.754 (16.992)$$

$$\frac{s}{\sqrt{96}} = 1.710$$

$$\frac{s}{\sqrt{190}} = 1.215$$

Result:

Azimuth deviation

$\mu \approx \bar{x} = 8.1$ cm with standard error 1.2 cm

$\sigma \approx s = 11.6$ cm with standard error 0.8 cm.

Height deviation

$\mu \approx \bar{x} = -11.6$ cm with standard error 1.7 cm

$\sigma \approx s = 16.8$ cm with standard error 1.2 cm.

For comparison we have for $\bar{x}$ and s also calculated the exact values (the numbers in parenthesis) from (11.7.3) and (11.7.4). It will be seen that the errors introduced by the grouping are small compared to the statistical fluctuations estimated in the standard errors. By calculating the expected theoretical distributions in the example, §10.5, we have used the exact values, but the approximate values would give exactly the same graphs.

*§11.13. **The correlation coefficient.** If we have a simultaneous measurement of two not necessarily independent physical quantities, we base the theory of errors on the two-dimensional normal distribution (cf. §7.6), in which we shall now write x for t and y for u. Applying the method of maximum likelihood to a sample consisting of n pairs of measurements, $(x_1, y_1), \cdots, (x_n, y_n)$, we again obtain as estimates of μ_x, μ_y, σ_x and σ_y, again substituting $n-1$ for n in the estimates of the dispersions

$$\mu_x \approx \bar{x} = \frac{[x]}{n}, \qquad \mu_y \approx \bar{y} = \frac{[y]}{n} \tag{1}$$

$$\sigma_x \approx s_x = \sqrt{\frac{[v_x v_x]}{n-1}}, \qquad \sigma_y \approx s_y = \sqrt{\frac{[v_y v_y]}{n-1}} \tag{2}$$

$$v_{x_i} = x_i - \bar{x}, \qquad v_{y_i} = y_i - \bar{y}.$$

For the correlation coefficient we find

$$\rho \approx r = \frac{[(x-\bar{x})(y-\bar{y})]}{(n-1)s_x s_y} = \frac{[v_x v_y]}{\sqrt{[v_x v_x][v_y v_y]}}. \tag{3}$$

Exercise. Verify (1)–(3), and find the corresponding standard errors. However, for ρ the estimate r is not normally distributed unless n is a very large number, so we should be careful when applying its standard error as if r were normal.

For the numerical computation of $[v_x v_y]$ we have the following formulae corresponding to the previous formulae for $[vv]$:

$$[v_x v_y] = [v_x{}^0 v_y{}^0] - \frac{[v_x{}^0][v_y{}^0]}{n} \tag{4}$$

$$[v_x v_y] = [xy] - \frac{[x][y]}{n} \tag{5}$$

$$[v_x v_y] = \Delta t_x \, \Delta t_y \left([n_{xy} r_x r_y] - \frac{[n_x r_x][n_y r_y]}{n} \right) \tag{6}$$

(check). Here $v_{x_i}{}^0 = x_i - x^0$, $v_{y_i}{}^0 = y_i - y^0$, $t_{x_i} - t_x{}^0 = r_{x_i} \Delta t_x$, $t_{y_i} - t_y{}^0 = r_{y_i} \Delta t_y$, $n_{x_i} = \sum_j n_{x_i y_j}$, and $n_{y_j} = \sum_i n_{x_i y_j}$, with (6) referring to a **two dimensional grouping** (cf. § 10.5): we divide a suitable interval on the x-axis into $m_x (< n)$, not necessarily equal, subintervals Δt_{x_i} and one on the y-axis into $m_y (< n)$, not necessarily equal, subintervals Δt_{y_j}. In the xy-plane we thus obtain $m_x m_y$ rectangles, the middle points of which we denote by (t_{x_i}, t_{y_j}). For each rectangle we count the number of measurements, $n_{x_i y_j}$, for which the results lie in the rectangle

$$t_{x_i} - \frac{\Delta t_{x_i}}{2} < x \leqq t_{x_i} + \frac{\Delta t_{x_i}}{2}, \qquad t_{y_j} - \frac{\Delta t_{y_j}}{2} < y \leqq t_{y_j} + \frac{\Delta t_{y_j}}{2}. \tag{7}$$

Example. For the 96 shots of the example, § 10.5, we have the following values of $n_{x_i y_j}$:

Height deviation y \ Azimuth deviation x	−30	−20	−10	0	10	20	30	40	50	Total: n_y
−60	0	0	0	0	0	0	0	0	0	0
−50	0	0	0	1	0	2	0	0	0	3
−40	0	0	1	1	1	2	0	0	0	5
−30	0	1	1	3	5	2	1	0	0	13
−20	0	1	3	7	3	2	2	0	0	18
−10	0	0	2	6	10	3	0	0	0	21
0	0	0	1	6	6	6	1	1	0	21
10	0	0	0	3	3	3	1	0	0	10
20	0	0	1	1	2	1	0	0	0	5
30	0	0	0	0	0	0	0	0	0	0
Total: n_x	0	2	9	28	30	21	5	1	0	96

Such a table is called a **correlation table.** The two distributions in the margin are just the **empirical marginal distributions** (giving the frequencies of one variable independent of the values of the other,

cf. § 4.13). We see that they agree with the two distributions of the example, § 11.12. In order to compute $[n_{xy}r_x r_y]$ we first compute $\sum_j n_{x_i y_j} r_{y_j} = [n_{x_i y} r_y]$ for each *fixed* x_i, multiply the result by r_{x_i}, and sum. The computations then run as follows, using the results of the example, § 11.12:

r_x	$[n_{x_i y} r_y]$	$[n_{x_i y_j} r_y] r_{x_i}$
−3	0	0
−2	−5	10
−1	−13	13
0	−33	0
1	−28	−28
2	−26	−52
3	−6	−18
4	0	0
5	0	0
	−111	−75

$$[n_{xy}r_x r_y] = -75$$

$$[v_x v_y] = 100\left(-75 + \frac{78 \cdot 111}{96}\right) = 1518.7$$

$$r = \frac{1518.7}{\sqrt{12863 \cdot 26666}} = 0.0820.$$

We note that the sum of the figures in the second column gives us a check since

$$\sum_i \sum_j n_{x_i y_j} r_{y_j} = \sum_j \left(\sum_i n_{x_i y_j}\right) r_{y_j} = \sum_j n_{y_j} r_{y_j} = [n_y r_y] = -111$$

from the example, § 11.12. In practice the whole computation is, of course, carried out in one single scheme, the schemes of the example, § 11.12, being added to those of this example.

§ 11.14. In some *statistical analyses* the problem is to decide whether the true value of a normally distributed variable is 0. In order to test whether the estimate, $\bar{x}$, of μ is compatible with the hypothesis $\mu = 0$, we form the new random variable

$$t = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}, \tag{1}$$

which, as shown in Exercise 2, § 7.9, has the t-distribution (7.9.3) with $f = n - 1$. From Table III we can then test whether the value of t observed lies within reasonable limits. In practice the 5% limit is usually chosen, which means that $\bar{x}$ is considered as deviating **significantly** from 0, i.e., more than can be expected from the statistical fluctuations, if the corresponding t-value lies outside the interval for which $P(|t| \geqq t) = 5\%$.

Example. Considering the 96 shots in the example, § 10.5, we may thus ask whether the deviations of the striking points from the aim point (0, 0) are so large as to indicate systematic errors in the aim mechanism of the machine gun, or whether they are only what must be expected from the statistical fluctuations. From the values given in the example, § 10.5, we get for the azimuth deviation $t_x = 8.125/1.188 = 6.84$, and for the height deviation $t_y = -11.562/1.710 = -6.76$. Since $f = 96 - 1 = 95$ it follows from Table III that these t-values are falling far outside the 5% limit of 1.986. Furthermore, for this value of f, t is very nearly normally distributed; i.e., Table II may be applied showing that the t-values formed are even outside the $P = 10^{-9}$ limit. Consequently the deviations are highly significant, and thus the series of shots mentioned requires a closer investigation of the causes of the error.

§ **11.15.** In other *statistical analyses* the conditions of the experiments are varied in order to investigate whether or not certain factors have any influence (cf. § 9.5). Performing respectively n_1 and n_2 equally accurate measurements before and after the conditions are varied we obtain two estimates, $\bar{x}_1$ and $\bar{x}_2$, of the true value. The problem is then to test whether or not the estimates are compatible with the hypothesis that the true value is the same, i.e., whether or not their difference is larger than can be expected from the statistical fluctuations. To test this we form the random variable

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}, \qquad s = \sqrt{\frac{f_1 s_1^2 + f_2 s_2^2}{f_1 + f_2}}, \tag{1}$$

$$f_1 = n_1 - 1, \qquad f_2 = n_2 - 1,$$

which, as shown in Exercise 3, § 7.9, has the t-distribution (7.9.3) with $f = f_1 + f_2$. (We note that as shown in § 11.10 s is the best estimate of σ we can obtain from the combined $n_1 + n_2$ measurements.) By means of Table III we can then test whether the value of t found lies within reasonable limits. In practice it is customary to choose the 5% limit and consider the difference between $\bar{x}_1$ and $\bar{x}_2$ as **significant**

if the corresponding t-value lies outside these limits. In this case we do not consider $\bar{x}_1$ and $\bar{x}_2$ as estimates of the same true value, or, in other words, *the larger the value of t the more the factor in question has had a significant effect.*

Example. In the measurement of some spectral lines the problem is whether or not we are observing a new line. Each measurement is repeated 4 times. In the first set the following values have been found:

$$50.339 \quad 50.340 \quad 50.337 \quad \text{and} \quad 50.339,$$

whence

$$\bar{x}_1 = 50.33875 \sim 50.339 \qquad \text{and} \qquad s_1 = 0.00126 \sim 0.001;$$

in the second set the values

$$50.339 \quad 50.330 \quad 50.333 \quad \text{and} \quad 50.333,$$

whence

$$\bar{x}_2 = 50.33375 \sim 50.334 \qquad \text{and} \qquad s_2 = 0.00378 \sim 0.004.$$

Thus we get $s = \sqrt{\dfrac{3{s_1}^2 + 3{s_2}^2}{6}} = 0.00281$ and $t = \dfrac{0.00500}{0.00281}\sqrt{2} = 2.52$. Since $f = 4 + 4 - 2 = 6$ it is seen from Table III that the value of t falls slightly outside the 5% limit, 2.447, but far within the 1% limit, 3.707. Thus $\bar{x}_1$ and $\bar{x}_2$ do differ significantly, and consequently we must regard them as belonging to different spectral lines, although we cannot ascribe great certainty to this conclusion.

§11.16. A too-large value of t in the preceding topic may also be attributable to the fact that the two sets of measurements are *not* equally accurate, i.e., they belong to two different values of σ. However, this may be tested separately. In many *statistical analyses* it is of importance to test whether or not the difference between two estimates of σ, found by two different methods of measurement, is **significant**, i.e., larger than can be expected from the statistical fluctuations; in other words, *whether or not the two measurements are equally accurate*, i.e., have the same parameter σ. Let s_1 and s_2 be the two estimates of σ based on n_1 and n_2 measurements respectively. We then form the new random variable

$$w^2 = \left(\frac{s_1}{s_2}\right)^2, \qquad s_1 \geqq s_2, \tag{1}$$

which, as shown in §7.11, has the distribution (7.11.4) with $f_1 = n_1 - 1, f_2 = n_2 - 1$. From Table VI we can then test whether or not the value of w^2 observed lies within reasonable limits. In practice the

5% limit is chosen, $P(w^2 \geqq w^2) = 5\%$, and, if w^2 is larger than this limit, the difference between s_1 and s_2 is considered **significant,** i.e., the two methods of measurements are not considered equally accurate.

Example 1. In the measurement of spectral lines mentioned in the example, § 11.15, we found $s_1 = 0.00126$ and $s_2 = 0.00378$, i.e., $w^2 = \left(\frac{0.00378}{0.00126}\right)^2 = 9.00$. Since $f_1 = f_2 = 4 - 1 = 3$ we see from Table V that this value falls just within the 5% limit 9.28. Thus the difference between s_1 and s_2 is not significant, and consequently there is no reason for suspecting that the two sets of measurements are not equally accurate.

Example 2. For the 96 shots of the example, § 10.5, we found $s_1 = 16.992$ and $s_2 = 11.393$, whence $w^2 = (16.992/11.393)^2 = 2.23$. From Table V we see that for $f_1 = f_2 = 96 - 1 = 95$ the 5% limit is smaller than 1.70. Thus the difference is significant and the dispersion for the azimuth deviation and for the height deviation cannot be regarded as being equal.

§ 11.17. In the practical application of the theory of errors it is important that measurements encumbered with coarse errors be rejected since *even a single coarse error can completely falsify the result of a series of measurements.* However, having based the theory of errors on the normal distribution which in principle allows the occurrence of arbitrarily large errors, although with negligible probabilities (cf. Table II), the problem is how to distinguish between *coarse* errors and particularly large *random* errors. *The only certain method is to reject during the measurements themselves suspicious measurements,* when certain circumstances seem to indicate that the conditions of the observations have not been constant (vibrations, sudden changes of temperature, and so forth). In such cases a closer investigation is necessary to decide whether or not the measurement has to be rejected. However, in practice such an investigation can be difficult or even impossible, and, since, as a rule, a coarse error is conspicuous by its considerable deviation from the other measurements, *it is tempting to let the magnitude of the error be the sole deciding criterion for the rejection of coarse errors,* especially when the measurements are performed by less skilled observers.

In the course of time many different rules for rejecting coarse errors have been proposed, especially in ordnance.[1] If the parameters μ and

[1] We then speak of rules for rejecting stragglers, German, **Ausreisserregeln.** For a more detailed discussion of such rules see, e.g., Arley, *Danske Vid. Selsk. Mat.-fys. Medd.*, Vol. XVIII, No. 3, 1940, Chapter IV; or Cranz, *Lehrbuch der Ballistik,* Bd. I, p. 420 ff.

σ were known we could take a suitable tolerance limit, e.g., the 0.1% limits $\mu \pm 3.29\sigma$. However, as a rule, only the estimates $\bar{x}$ and s are available, and thus one natural procedure among others is to consider the relative deviations

$$r_i = \frac{x_i - \bar{x}}{\sqrt{\frac{n-1}{n}}\, s}, \tag{1}$$

of which, as shown in §7.10, each has as marginal distribution the r-distribution (7.10.2) with $f = n - 2$. For large values of n this distribution is practically normal, so the r-limits will approach those of the normal distribution. By means of Table IV we can test whether or not the relative deviations lie within reasonable limits. Since the rejection of measurements for which the r-value lies outside certain limits means in reality that a distribution other than the normal is taken as the basis of the theory of errors, such limits must correspond to a very small probability in order that the change be so small as to be negligible. As a suitable limit we may choose the 0.1% limit, $P(|r| \geqq r) = 0.1\%$, and consequently reject a measurement if its r-value is outside these limits.

Example. For the 10 measurements of the example, §11.11, we get the 10 relative deviations

0.563	0.295
1.100	−0.241
−1.315	−0.778
0.027	0.832
1.368	−1.851.

Since $f = 10 - 2 = 8$, it is seen from Table IV that these r-values all lie within the 0.1% limit, 2.616, and even within the 5% limit, 1.895. Thus there is not the slightest reason for suspecting that any of these measurements is encumbered with coarse errors.

We emphasize that any schematic rule for rejecting suspected coarse errors must be applied with a certain caution, especially for small values of n.

Exercise. Show that for the 4 measurements 21,790, 21,789, 21,789, 21,789 the first relative deviation is equal to the maximum value of r, viz., $\sqrt{3}$. In spite of this fact no one would of course reject the first measurement.

*§11.18. An important form of *statistical analysis* is the investigation of dependence or *correlation* between certain phenomena (cf. §9.5). If it may be assumed that the joint distribution function of

x and y is normal, the problem is to test the hypothesis that the true correlation coefficient $\rho\{x, y\}$ is zero. To test this we calculate the best estimate of ρ, the empirical correlation coefficient r_{xy} given in (11.13.3), and the new random variable

$$r = \sqrt{n-1}\, r_{xy}. \tag{1}$$

It may be shown that r has for $\rho = 0$ the r-distribution (7.10.2) with $f = n - 2$.[1] By means of Table IV we may test whether or not r lies within reasonable limits. Taking as such limits the 5% limits, $P(|r| \geqq r) = 5\%$, this means that we consider that the value of r observed deviates **significantly** from 0 if it falls outside these limits or, in other words, that the hypothesis $\rho = 0$ is incompatible with the sample measured, i.e., that there is a correlation between x and y.

Example. For the 96 shots of the example, § 10.5, we have in the example, § 11.13, found the estimate $r = 0.0820$ for the correlation coefficient between the azimuth and height deviation of the machine gun. Thus $r = \sqrt{95} \cdot 0.0820 = 0.799$. Since $f = 96 - 2 = 94$, it is seen from Table IV that the r-value here is far inside the 5% limit of 1.956. Furthermore, since the r-distribution for such large values of f is very nearly normal, Table II may be applied, showing that this r-value falls even within the 40% limit of 0.8416. Thus r does not deviate significantly from 0, and consequently there is no correlation between the azimuth and height deviation for the machine gun investigated.

§ 11.19. We emphasize that the common characteristic features of the methods described in § 11.14–§ 11.18 are, first, that the methods do not presuppose any knowledge of the true values of the parameters which in practice are known only rarely; second, that they take into account the statistical fluctuations which are due to the fact that in practice we must often be content with *small* samples, consisting of only a small number of observations; and third, that they give us complete control of the certainty of our conclusions. However, it should also be stressed that the various tests discussed allow us to draw only *negative* conclusions. Thus if the hypothesis is that a true value is zero and the observations give a t-value which has only a small probability, then it is obvious that the hypothesis does not agree very well with the observations. But if we consider two hypotheses regarding the true values, μ_1 and μ_2, and the observations give t-values for these

[1] See, e.g., Cramér, *Mathematical Methods of Statistics*, § 29.7. A table giving directly the distribution of r_{xy} is found in Fisher and Yates, *Statistical Tables*, Table VI.

two hypotheses within say the 90% and the 70% limits respectively, we cannot conclude that μ_1 is a better hypothesis than μ_2 but only that they are both compatible with the observations. Furthermore, it should also be stressed again that it is quite arbitrary when we choose the 5% level for discarding unlikely hypotheses. As discussed in § 9.3, it is always, not only here, a question of personal choice where to make the distinction between agreement and disagreement of theory and observations. The only justification for the value 5% is that in practice this has turned out to be a suitable value: on the one hand it is sufficiently large for actually discarding "false" hypotheses, on the other it is sufficiently small for discarding only a few "true" hypotheses giving rise in a random way to large deviations.

If, however, the tables used in the various tests are not at our disposal, the tests may be performed roughly by neglecting the statistical fluctuations of s. Thus if we want to find out whether or not two averages agree we can do it roughly by computing the dispersion of the difference and testing the significance of the difference by means of the tolerance limits of the normal distribution. If the difference is many times as large as its dispersion it is certainly significant, but, if it is of the same order of magnitude and the number of observations is small, as, e.g., in the example, § 11.15, certain conclusions can be drawn only by means of the t-table, since in such cases the statistical fluctuations of s cannot be neglected.

§ 11.20. For the practical applications of the theory of errors the question of how many observations should be made is of great importance. Of course this depends on the circumstances in question, especially on the knowledge we may have in advance as to the statistical fluctuations expected. *In accurate measurements we should demand that the dispersion of $\bar{x}$ be small compared with $\bar{x}$ and that the dispersion of s be small compared with s.* In general, results based on 10 measurements are considered reliable, and more than 20 measurements are seldom used for economical reasons, as regards both time and expenses.

Since the dispersion of $\bar{x}$ tends to 0 when the number of observations, n, is increased indefinitely it might in principle be thought possible to increase the accuracy in the determination of the true value indefinitely by increasing the number of measurements, rather than using more precise measuring instruments. However, this belief is erroneous although it is often met.

Example. In the measurement of the distance between two marks on a steel rod, mentioned in § 11.3, we get exactly the dispersion 0 for $\bar{x}$, but hardly anyone would conclude from this that the true length is exactly 2 m. If we measured next the same length with

a rule with divisions of only centimeters the dispersion of $\bar{x}$ might for a sufficiently large number of measurements be of the order of magnitude of 10^{-4} m. However, hardly anyone would expect the length thus obtained to agree with the value 2.0078 m found by measuring the length by means of a much finer measuring instrument.

It should be remembered, first, that $\bar{x}$ and s are only estimates of the parameters in the theoretical distribution by means of which we describe our observations, and, although the dispersions of these estimates tend to 0 for $n \to \infty$, s itself will *not* tend to 0 but converge in probability to a constant, σ, different from 0. Second, it should be remembered that it is quite arbitrary that we *define* the true value as being exactly the mean value, μ, of our distribution function. Any value, say within $\mu \pm \sigma$, might with equal validity be called the "true" value. Third, it should be remembered that from this definition the true value depends also on the method of measurement. Thus in the example we are dealing with three different methods of measurement; i.e., we have three different random variables each with a distribution function of its own. Finally we must remember that using precisely the arithmetic average as the best estimate of the parameter μ is founded upon the hypothesis that the normal distribution may be taken as the foundation of the theory of errors. But this hypothesis can be verified only to a certain degree (cf. the next topic) since any *empirically* found distribution is discontinuous and, therefore, can be described only to a certain approximation by means of a continuous or, in particular a normal distribution when the variations in the values measured are large compared to the smallest unit of the measuring instrument.

For all these reasons it is, in general, meaningless to write the estimates $\bar{x}$ and s to more than one place in addition to the number of places in the separate measurements.

§ 11.21. In this topic we shall show how in practice one may test the fundamental hypothesis of the theory of errors, i.e., that directly measured quantities may be satisfactorily described by means of the normal distribution.

The simplest procedure is to form the histogram and compare it with the normal probability density $\frac{1}{\sigma}\psi\left(\frac{x-\mu}{\sigma}\right)$ by inserting $\bar{x}$ for μ and s for σ and computing the density by means of Table I. In the example, § 10.5, we have given an example of this procedure.

Another, and better, procedure is to form the sum polygon and to compare it with the normal distribution function $\Psi\left(\frac{x-\mu}{\sigma}\right)$, inserting

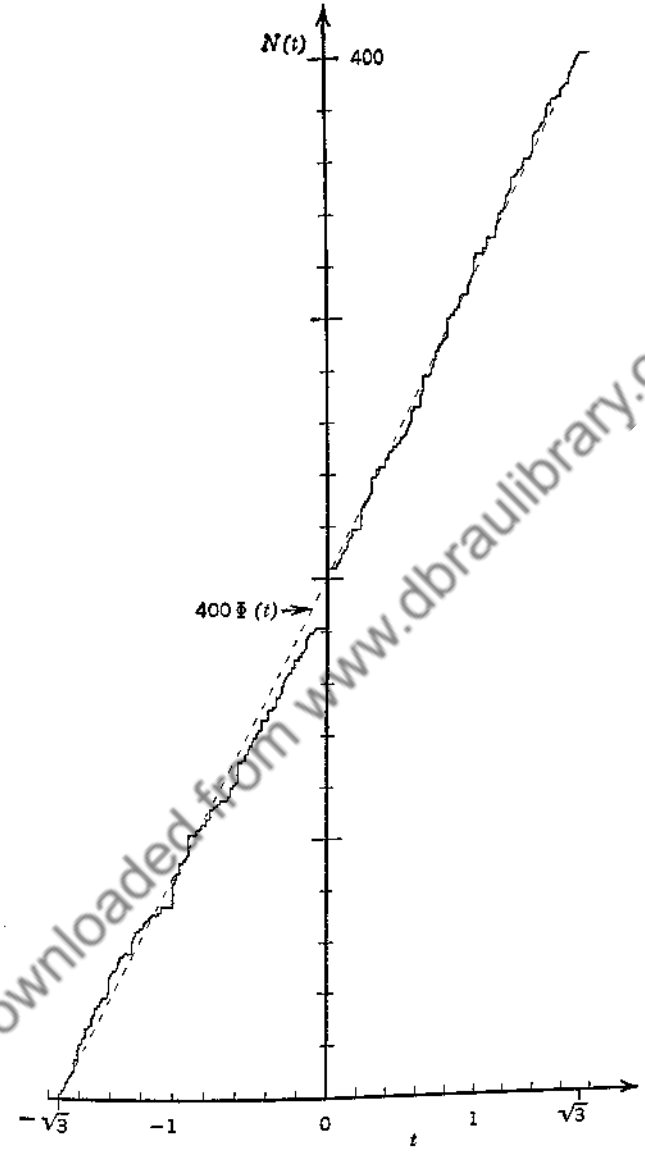

FIG. 19.

$\bar{x}$ for μ and s for σ. In the example, § 10.6, we have demonstrated this procedure.

A still better procedure is to use the probit diagram method discussed in § 10.7, whereby the graph of the normal distribution function is transformed into a straight line. In Example 1, § 10.7, we have given an example of this procedure.

As a rule, in practice we have only small samples consisting of a small number of measurements. In such cases the methods mentioned fail since these methods are based on a large number of observations. However, it often happens that we have a large number of small samples which belong to different values of both μ and σ but in which the quantities measured and the methods of measurement applied are of the same nature, so that the distribution functions may safely be assumed to be of the same type, viz., by hypothesis the normal. We can then form all the relative errors (11.17.1) and compare their distribution with the theoretical r-distribution given in (7.10.2) and tabulated in Table IV.

Example 1. In Fig. 19 is shown the sum polygon, $400F(t) = N(t)$, for 400 relative errors computed from 100 samples each consisting of 4 measurements of spectral lines.[1] For comparison we have also plotted the corresponding theoretical curve, $400\,\Phi(t)$, which from (7.10.2) for $f = 4 - 2 = 2$ reduces to a straight line between the points $(-\sqrt{3}, 0)$ and $(\sqrt{3}, 400)$, which means that the relative errors are for $f = 2$ uniformly distributed, each value between $-\sqrt{3}$ and $\sqrt{3}$ being equally probable. It is seen that the agreement is very good, although in this case the variations are so small that they are not large compared to the smallest unit, a condition usually necessary in approximating an empirical distribution by a continuous distribution function.

It should be remarked that strictly speaking we should consider the distributions of the relative errors, Nos. 1, 2, 3, and 4, separately, since otherwise they are not independent (this is done in Arley[1]). However, when the number of samples is large this dependency may be neglected as proved in Arley.[1]

Example 2. Finally we can also form the relative errors of a *single* sample and compare their sum polygon with the normalized, normal distribution function $\Psi(t)$. However, since the r-distribution is only approximately normal for large values of n and since the relative errors within a single sample are dependent, such a comparison can give only a rough test of whether or not the quantity considered is normally distributed.[1] In Fig. 20 we have drawn the sum polygon of the 10 measurements of the example, §11.11, and for comparison $\Psi(t)$. It will be seen that the agreement is not bad.

§11.22. So far we have considered only direct measurements; but as mentioned in §11.5 most physical measurements are *indirect*, i.e., the results are obtained only indirectly from the observations, the

[1] This example is taken from Arley, *Danske Vid. Selsk. Mat.-fys. Medd.*, Vol. XVIII, No. 3, 1940.

quantity z under consideration being a function of other quantities $x_1, x_2, \cdots, x_\nu$, which are measured directly. Assuming $x_1, \cdots, x_\nu$ to be independent and the conditions mentioned in §6.5 to be fulfilled the mean and the dispersion of z are to a good approximation given by (6.5.9) and (6.5.10). Inserting herein estimates $\mu_1 \approx \bar{x}_1, \cdots, \mu_\nu \approx \bar{x}_\nu, \sigma_1 \approx s_1, \cdots, \sigma_\nu \approx s_\nu$ we obtain

$$\mathfrak{M}\{z\} \approx \bar{z} \sim f(\bar{x}_1, \cdots, \bar{x}_\nu) \tag{1}$$

$$\sigma^2\{z\} \approx s^2 \sim \left(\frac{\partial f}{\partial x_1}\right)^2 s_1^2 + \cdots + \left(\frac{\partial f}{\partial x_\nu}\right)^2 s_\nu^2, \tag{2}$$

in which the partial derivatives have to be taken in the point $(\bar{x}_1, \cdots, \bar{x}_\nu)$.

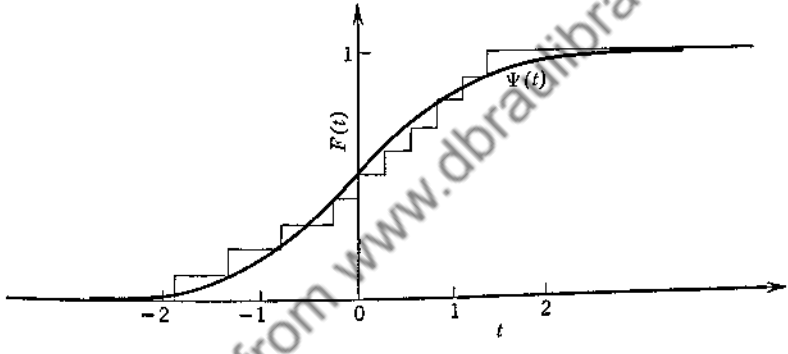

FIG. 20.

In many cases, occurring especially in physics and technology, z is a so-called *logarithmic* function, i.e.,

$$z = c \cdot x_1^{k_1} \cdot x_2^{k_2} \cdot \cdots \cdot x_\nu^{k_\nu}, \tag{3}$$

where $k_1, \cdots, k_\nu$ are positive or negative constants and c is also a constant. From (6.5.9) and (6.5.10) we then find

$$\sigma^2\{z\} \sim \mathfrak{M}^2\{z\} \left(k_1^2 \left(\frac{\sigma_1}{\mu_1}\right)^2 + \cdots + k_\nu^2 \left(\frac{\sigma_\nu}{\mu_\nu}\right)^2\right), \tag{4}$$

which may also be written

$$\left(\frac{\sigma_z}{\mu_z}\right)^2 \sim k_1^2 \left(\frac{\sigma_1}{\mu_1}\right)^2 + \cdots + k_\nu^2 \left(\frac{\sigma_\nu}{\mu_\nu}\right)^2. \tag{5}$$

Exercise. Verify these formulae. Also, if $k_1 \frac{\sigma_1}{\mu_1} \sim k_2 \frac{\sigma_1}{\mu_2} \sim \cdots \sim k_\nu \frac{\sigma_\nu}{\mu_\nu}$, show that (5) may be approximated by

$$\frac{\sigma_z}{\mu_z} \sim \frac{1}{\sqrt{\nu}} \left(k_1 \frac{\sigma_1}{\mu_1} + \cdots + k_\nu \frac{\sigma_\nu}{\mu_\nu} \right). \tag{6}$$

This formula is frequently used, especially by engineers; it always gives a value smaller than (5) except where the single contributions are exactly equal (check). Therefore, since (5) is just as easy to apply, (5) should always be used.

Example 1. Let us consider the measurement of specific resistance $\rho = \dfrac{\pi R d^2}{4l}$, and let us assume that the total resistance R, the diameter d, and the length l have been measured with the relative dispersions 2%, 3%, and 3% respectively. From (5) we find

$$\left(\frac{\sigma_\rho}{\mu_\rho}\right)^2 \sim \left(\frac{\sigma_R}{\mu_R}\right)^2 + 4\left(\frac{\sigma_d}{\mu_d}\right)^2 + \left(\frac{\sigma_l}{\mu_l}\right)^2 = 49,$$

i.e.,

$$\frac{\sigma_\rho}{\mu_\rho} \sim 7\%.$$

(From (6) we would have found $\sigma_\rho/\mu_\rho \sim 11/\sqrt{3}\% = 6\%$.)

Formula (5) is often used for planning the measurement. Thus if we demand that the relative dispersion of z be less than a given value and if, furthermore, we demand that this tolerance be uniformly distributed, i.e., that the separate measurements shall contribute the same amount to the relative dispersion of z, we can find the relative dispersions with which we have to perform the direct measurements of $x_1, \cdots, x_\nu$.

Example 2. Thus, if in Example 1 we demand that the relative dispersion of ρ be at most 1%, we find

$$\left(\frac{\sigma_R}{\mu_R}\right)^2 < \tfrac{1}{3} \cdot 0.01^2, \quad \text{i.e.,} \quad \frac{\sigma_R}{\mu_R} < 0.6\%$$

$$4\left(\frac{\sigma_d}{\mu_d}\right)^2 < \tfrac{1}{3} \cdot 0.01^2, \quad \text{i.e.,} \quad \frac{\sigma_d}{\mu_d} < 0.3\%$$

$$\left(\frac{\sigma_l}{\mu_l}\right)^2 < \tfrac{1}{3} \cdot 0.01^2, \quad \text{i.e.,} \quad \frac{\sigma_l}{\mu_l} < 0.6\%.$$

It should be remarked that in practice this rule of the uniform distribution of the tolerance need not always be appropriate to use, since we have to plan our measurements with due regard to our knowledge of the possibility of making the dispersions of the various direct measurements small.

12.

APPLICATION OF THE THEORY OF PROBABILITY TO THE THEORY OF ADJUSTMENT

§12.1. In the theory of errors, discussed in Chapter 11, we have considered the simple case of n independent measurements of one and the same physical quantity. In the **theory of adjustment**[1] we consider the more general problem of n independent normally distributed measurements of *different* quantities which are not free but which, as we know beforehand, are subjected to certain relations. Thus if in a problem of surveying we consider three points A, B, and C marked out in the field, and if we measure the three angles of the triangle ABC, their sum must be equal to 180°. Now, as we have discussed in §11.3, the result of any physical measurement is a random variable, and consequently the sum mentioned will in general be found to deviate somewhat from 180°.

Let us consider the general case in which we have performed n independent measurements, the results of which we shall denote $l_1, \cdots, l_n$, of n physical quantities, $\boldsymbol{l}_1, \cdots, \boldsymbol{l}_n$, the true values, i.e., the mean values, of which, $\lambda_1, \cdots, \lambda_n$, satisfy $r < n$ equations,[2]

$$F_j(\lambda_1, \cdots, \lambda_n) = 0, \qquad j = 1, 2, \cdots, r, \tag{1}$$

called the **fundamental equations.** If for $\lambda_1, \cdots, \lambda_n$ we insert the observed values, $l_1, \cdots, l_n$, these equations will, in general, not be satisfied. It is the purpose of the theory of adjustment to obtain from the measured values the best estimates, $\bar{l}_1, \cdots, \bar{l}_n$, of the parameters $\lambda_1, \cdots, \lambda_n$, which like the λ's also satisfy the fundamental equations.[3] We solve this problem by means of the maximum

[1] French, **théorie de l'ajustement;** German, **Ausgleichungstheorie.**

[2] Cf. footnote 2, p. 137. In this chapter because of the large number of variables occurring it will be more convenient to denote the mean values by the corresponding Greek letters and their estimates by the corresponding Roman letters with a bar above, since these estimates will turn out to be simple generalizations of the arithmetic average values, viz., weighted average values (cf. §6.4).

[3] The solution of this problem is denoted by a somewhat unfortunate term "adjustment," which may give rise to the misunderstanding that we have to "change" the measured values a little in order to make them satisfy (1). Thus the term "theory of adjustment," now commonly used, is unfortunate.

likelihood method (§ 10.10). If the dispersion of l_i is called σ_i, the likelihood function (10.10.3) becomes

$$P(l_1, \cdots, l_n; \lambda_1, \cdots, \lambda_n, \sigma_1, \cdots, \sigma_n) =$$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma_1 \cdots \sigma_n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(l_i-\lambda_i)^2}{\sigma_i^2}\right] =$$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma_1 \cdots \sigma_n} \exp\left[-\frac{1}{2}\left[\frac{1}{\sigma^2}\epsilon\epsilon\right]\right], \quad (2)$$

where

$$\epsilon_i = l_i - \lambda_i, \qquad i = 1, 2, \cdots, n, \quad (3)$$

are called the "true" errors of the system and [ ] is the Gaussian symbol for a sum (11.5.2) commonly used in the theories of errors and of adjustment. Since in practice the ratios between the σ_i's are, as a rule, known, it is convenient to introduce n positive numbers $p_1, \cdots, p_n$ and a proportionality factor σ defined by

$$p_1\sigma_1^2 = p_2\sigma_2^2 = \cdots = p_n\sigma_n^2 = \sigma^2. \quad (4)$$

The p_i's are called the **weights** of the n measurements, and the (arbitrary) proportionality factor σ is called the **dispersion** (or the mean error of an observation of unit weight). Since $\sigma^2\left\{\frac{[x]}{n}\right\} = \frac{\sigma^2\{x\}}{n}$, the weight of an arithmetic average of n equally accurate measurements is n times as large as the weight of a single measurement. This fact is the reason for the notation "weights," because the arithmetic average contains all n measurements and has, therefore, the same weight as the n measurements together (cf. Example 1, § 12.9). Furthermore, in practice, this is often how the weights occur, since each quantity, l_i, is measured several times, n_i, and the average is then inserted into the equations as a directly measured value l_i with the weight n_i.

Introducing the weights into (2) we get

$$P(l_1, \cdots, l_n; \lambda_1, \cdots, \lambda_n, \sigma) =$$

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n (p_1 \cdots p_n)^{1/2} \exp\left[-\frac{1}{2\sigma^2}[p\epsilon\epsilon]\right]. \quad (5)$$

From (10.10.5) the best estimates $\bar{l}_1, \cdots, \bar{l}_n$ of $\lambda_1, \cdots, \lambda_n$ are the values which when inserted in (5) maximize P for fixed values of $l_1, \cdots, l_n$ and at the same time satisfy the fundamental equations (1).

Now, since $\lambda_1, \cdots, \lambda_n$ occur only through $[p\epsilon\epsilon]$ this is equivalent to the condition that $[p\epsilon\epsilon]$ be a minimum. The errors

$$v_i = \bar{l}_i - l_i, \qquad i = 1, 2, \cdots, n \tag{6}$$

which minimize $[p\epsilon\epsilon]$ are called the **best corrections** of the system,[1] and we have

$$[pvv] = p_1(\bar{l}_1 - l_1)^2 + \cdots + p_n(\bar{l}_n - l_n)^2 = \text{minimum}, \tag{7}$$

where $[pvv]$ is called the **weighted square error sum.** This principle is called the **method of least squares,** being a generalization of (11.5.7). We stress that in (7) $l_1, \cdots, l_n, p_1, \cdots, p_n$ are known quantities and $\bar{l}_1, \cdots, \bar{l}_n$ are the unknowns.

§12.2. We distinguish two methods of adjustment, *adjustment by elements* and *adjustment by correlates* according to the form of the fundamental equations.

In *adjustment by elements*[2] the fundamental equations (12.1.1) are given in *parameter form*, the true values $\lambda_1, \cdots, \lambda_n$ being given as functions of a certain number, $m < n$, of free parameters, $\xi_1, \cdots, \xi_m$, which determine the system uniquely and which are called its **elements:**

$$\lambda_i = f_i(\xi_1, \cdots, \xi_m), \quad i = 1, 2, \cdots, n \quad (n > m). \tag{1}$$

Thus, for instance, a plane triangle is determined by one side and two angles, or by the three sides, but not by the three angles. In general, the elements can be chosen in many different ways, and the set of elements to be preferred must be decided in each separate case. The only conditions are that the system is uniquely determined by the elements and that these are free, subject to no other constraints than to vary only within certain intervals, their region of definition. Thus in a plane triangle we cannot choose the three angles as elements, since, first, the triangle is not uniquely determined by its angles and, second, the angles are not free variables, being subject to the condition that their sum is 180°.

If we are to measure a given system, the problem is to find the best estimates $\bar{x}_1, \cdots, \bar{x}_m$ of the true values $\xi_1, \cdots, \xi_m$ of the m elements $x_1, \cdots, x_m$, i.e., the values which on the one hand satisfy the r fundamental equations (1) and on the other minimize $[pvv]$. It

[1] We note that $v_1, \cdots, v_n$, as defined in (6), are not the errors, but the corrections (cf. footnote 1, p. 156). It is more natural in the theory of adjustment to consider the corrections rather than the errors.

[2] French, **compensation d'observations indirecte;** German, **Ausgleichung vermittelnder Beobachtungen.**

is, however, not always adequate, or even possible, to measure the elements themselves directly, and we must therefore measure them indirectly by measuring other parts of our system, which are all functions of the elements. Measuring just as many quantities as the system has elements we obtain m equations for determining the m unknown elements which are thus determined mathematically. However, as discussed in § 11.5, one measurement of an unknown is not sufficient. Therefore, to decrease the influence of the statistical fluctuations and to estimate their magnitude we measure *more* quantities than there are unknowns; in the following we shall assume $n > m$. We say that the system is **overdetermined,** and the additional measurements are said to be **overcomplete.**

By **adjustment by correlates**[1] the fundamental equations (12.1.1) are given as $r < n$ unsolved equations between the λ's:

$$f_j(\lambda_1, \cdots, \lambda_n) = 0, \quad j = 1, 2, \cdots, r \quad (n > r). \tag{2}$$

Having here n unknowns subjected to r constraints we can in principle arbitrarily choose $n - r = m$ free λ's, consider these as elements, and express the $n - m = r$ remaining λ's as functions of the first λ's. Consequently, we can again solve the problem by means of adjustment by elements, but the idea of the adjustment by correlates is to avoid solving the equations (2) by operating directly with the unsolved equations. Conversely we can in principle eliminate the elements from (1) thus obtaining the fundamental equations on the form (2). *Thus adjustment by elements and adjustment by correlates are only two different methods for solving the same problem, giving, of course, the same results.* In practice mixed adjustments may be met in which some of the fundamental equations are in parameter form, while others are in the form of unsolved equations between the λ's. We shall here treat only the pure cases.[2]

Inserting $\bar{l}_1, \cdots, \bar{l}_n$, given in (12.1.6), for $\lambda_1, \cdots, \lambda_n$ and $\bar{x}_1, \cdots, \bar{x}_m$ for $\xi_1, \cdots, \xi_m$ in the fundamental equations (1) or (2) we get by adjustment by elements the equations

$$l_i + v_i = f_i(\bar{x}_1, \cdots, \bar{x}_m), \qquad i = 1, 2, \cdots, n, \tag{3}$$

i.e., n equations for the $n + m$ unknowns $v_1, \cdots, v_n$ and $\bar{x}_1, \cdots, \bar{x}_m$.

[1] French, **compensation d'observations conditionnelles;** German, **Ausgleichung bedingter Beobachtungen.**

[2] For a discussion of mixed adjustments we refer to the textbooks in the theory of adjustment mentioned in the list of references.

By adjustment by correlates we get the equations

$$f_j(l_1 + v_1, \cdots, l_n + v_n) = 0, \qquad j = 1, 2, \cdots, r, \tag{4}$$

i.e., r equations for the n unknowns $v_1, \cdots, v_n$. In both cases we get fewer equations than unknowns, i.e., $n - r = m$. The remaining m equations are given by the condition that $[pvv]$ = minimum, as we shall see shortly.

The equations (3) or (4) are called the **equations of condition,** and we shall here assume that they are *linear*, considering the non-linear case separately in § 12.13,

$$l_i + v_i = a_{i0} + a_{i1}\bar{x}_1 + a_{i2}\bar{x}_2 + \cdots + a_{im}\bar{x}_m, \qquad i = 1, 2, \cdots, n \tag{5}$$

or

$$a_{j0} + a_{j1}(l_1 + v_1) + a_{j2}(l_2 + v_2) + \cdots + a_{jn}(l_n + v_n) = 0, \qquad j = 1, 2, \cdots, r. \tag{6}$$

It is not only inconvenient to write down all these equations but also difficult to keep the survey clear in longer calculations. We therefore look for a way of writing them in a simpler and more concise form. We have the ideal tool for this purpose in the **matrix** symbolism outlined in Appendix 2.

ADJUSTMENT BY ELEMENTS

§12.3. We introduce the matrices

$$\underset{n1}{\boldsymbol{L}} = \begin{Bmatrix} l_1 \\ l_2 \\ \cdot \\ \cdot \\ \cdot \\ l_n \end{Bmatrix}, \quad \underset{n1}{\boldsymbol{V}} = \begin{Bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{Bmatrix}, \quad \underset{nm}{\boldsymbol{A}} = \begin{Bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{Bmatrix},$$

$$\underset{n1}{\boldsymbol{A}_0} = \begin{Bmatrix} a_{10} \\ a_{20} \\ \cdot \\ \cdot \\ \cdot \\ a_{n0} \end{Bmatrix}, \quad \underset{m1}{\bar{\boldsymbol{X}}} = \begin{Bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_m \end{Bmatrix}. \tag{1}$$

The equations of condition (12.2.5) may then be written

$$\underset{n1}{\bar{\boldsymbol{L}}} = \boldsymbol{L} + \boldsymbol{V} = \boldsymbol{A}_0 + \boldsymbol{A} \cdot \bar{\boldsymbol{X}}. \tag{2}$$

It is customary to write as an abbreviation

$$\underset{n1}{\boldsymbol{N}} = \boldsymbol{L} - \boldsymbol{A}_0, \tag{3}$$

and the equations of condition then have the form

$$\underset{n1}{\boldsymbol{V}} = -\boldsymbol{N} + \boldsymbol{A} \cdot \bar{\boldsymbol{X}}. \tag{4}$$

In (4) we have n equations for the $n + m$ unknowns $\boldsymbol{V}$ and $\bar{\boldsymbol{X}}$, $\boldsymbol{N}$ and $\boldsymbol{A}$ being given quantities. The m remaining equations we get from the condition that $[pvv]$ = minimum, since a necessary condition for this is that all the m partial derivatives of $[pvv]$ with respect to $\bar{x}_1, \cdots, \bar{x}_m$ be zero:

$$\frac{\partial}{\partial \bar{x}_j}[pvv] = 2\left(p_1 v_1 \frac{\partial v_1}{\partial \bar{x}_j} + \cdots + p_n v_n \frac{\partial v_n}{\partial \bar{x}_j}\right) = 0,$$

$$j = 1, 2, \cdots, m. \tag{5}$$

From (4) we see that

$$\left\{\frac{\partial v_r}{\partial \bar{x}_s}\right\} = \{a_{rs}\} = \boldsymbol{A}, \tag{6}$$

and, introducing the **weight matrix** as the diagonal matrix

$$\underset{nn}{\boldsymbol{P}} = \begin{Vmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & p_n \end{Vmatrix}, \tag{7}$$

the m equations (5) may be written

$$\boldsymbol{A}^* \cdot \boldsymbol{P} \cdot \boldsymbol{V} = 0. \tag{8}$$

Exercise 1. Verify this. Also show that if we introduce the m matrices

$$\underset{n1}{\boldsymbol{A}_1} = \begin{Vmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ \cdot \\ a_{n1} \end{Vmatrix}, \quad \underset{n1}{\boldsymbol{A}_2} = \begin{Vmatrix} a_{12} \\ a_{22} \\ \cdot \\ \cdot \\ \cdot \\ a_{n2} \end{Vmatrix}, \quad \cdots, \quad \underset{n1}{\boldsymbol{A}_m} = \begin{Vmatrix} a_{1m} \\ a_{2m} \\ \cdot \\ \cdot \\ \cdot \\ a_{nm} \end{Vmatrix} \tag{9}$$

the m equations (8) may be written

$$[pa_1 v] = [pa_2 v] = \cdots = [pa_m v] = 0. \tag{10}$$

Exercise 2. Show that $[pvv]$ may also be written

$$[pvv] = \boldsymbol{V}^* \cdot \boldsymbol{P} \cdot \boldsymbol{V}. \tag{11}$$

Inserting V from (4) into (8) we can eliminate V and obtain equations containing only $\bar{X}$:

$$A^* \cdot P \cdot A \cdot \bar{X} = A^* \cdot P \cdot N. \tag{12}$$

As an abbreviation we introduce the matrix

$$\underset{mn}{A^*} \cdot \underset{nn}{P} \cdot \underset{nm}{A} = \underset{mm}{B}, \tag{13}$$

which is *symmetric* because

$$B^* = (A^* \cdot P \cdot A)^* = A^* \cdot P^* \cdot A^{**} = A^* \cdot P \cdot A = B. \tag{14}$$

Thus we finally get

$$B \cdot \bar{X} = A^* \cdot P \cdot N. \tag{15}$$

These m equations for the m unknown $\bar{x}$'s are called the **normal equations** of the system.

Exercise 3. Show that by means of (9) we may write (15) as

$$[pa_j a_1]\bar{x}_1 + [pa_j a_2]\bar{x}_2 + \cdots + [pa_j a_m]\bar{x}_m = [pa_j n], \quad j = 1, 2, \cdots, m, \tag{16}$$

which is the usual form of the normal equations as introduced by Gauss. Sometimes the symbol (yz) is introduced for $[pyz]$.

From (15) we see that the method of least squares actually leads to a unique solution, obtained by solving (15) for $\bar{X}$, which is done by multiplying from the left with B^{-1}:

$$\underset{m1}{\bar{X}} = \underset{mn}{C} \cdot \underset{n1}{N}, \tag{17}$$

$$C = B^{-1} \cdot A^* \cdot P. \tag{18}$$

We have here assumed that the determinant $|B| \neq 0$, because if $|B| = 0$ it would follow from the theory of linear equations that we had certain relations between the $\bar{x}$'s in contradiction to our assumption that the elements are free variables. From $\bar{X}$ given in (17) we may finally obtain V from (4) and $\bar{L}$ from (2):

$$V = (-E + A \cdot C) \cdot N \tag{19}$$

$$\bar{L} = A_0 + A \cdot C \cdot N. \tag{20}$$

Exercise 4. Show that for $n = m$ these formulae reduce to

$$\bar{X} = A^{-1} \cdot N, \quad \bar{L} = L \quad \text{and} \quad V = 0. \tag{21}$$

§12.4. We still have to verify that V given in (19) actually minimizes $[pvv]$. Let V' be another set of corrections and X' the corresponding set of elements, i.e., from (12.3.4)

$$V' = -N + A \cdot X'. \tag{1}$$

Subtracting (12.3.4) from (1) we get

$$\boldsymbol{V}' - \boldsymbol{V} = \boldsymbol{A} \cdot (\boldsymbol{X}' - \bar{\boldsymbol{X}}). \tag{2}$$

Using (12.3.11) we then obtain as a generalization of (11.5.6), since $\boldsymbol{V}' = (\boldsymbol{V}' - \boldsymbol{V}) + \boldsymbol{V}$,

$$[pv'v'] = \boldsymbol{V}'^* \cdot \boldsymbol{P} \cdot \boldsymbol{V}' = \boldsymbol{V}^* \cdot \boldsymbol{P} \cdot \boldsymbol{V} + (\boldsymbol{V}' - \boldsymbol{V})^* \cdot \boldsymbol{P} \cdot (\boldsymbol{V}' - \boldsymbol{V}) + \boldsymbol{V} \cdot \boldsymbol{P} \cdot \boldsymbol{A} \cdot (\boldsymbol{X}' - \bar{\boldsymbol{X}}) + (\boldsymbol{X}' - \bar{\boldsymbol{X}})^* \cdot \boldsymbol{A}^* \cdot \boldsymbol{P} \cdot \boldsymbol{V} =$$

$$[pvv] + [p(v' - v)^2], \tag{3}$$

the third and fourth terms being zero owing to (12.3.8). But here the last expression is non-negative and only zero for $\boldsymbol{V}' = \boldsymbol{V}$. Thus $[pv'v']$ is a minimum for $\boldsymbol{V}' = \boldsymbol{V}$.

Exercise. Show that

$$[p\epsilon\epsilon] = q^2 + (\bar{\boldsymbol{X}} - \boldsymbol{\Xi})^* \cdot \boldsymbol{B} \cdot (\bar{\boldsymbol{X}} - \boldsymbol{\Xi}), \tag{4}$$

where

$$q = \sqrt{[pvv]}, \qquad \underset{m1}{\boldsymbol{\Xi}} = \begin{Bmatrix} \xi_1 \\ \xi_2 \\ \cdot \\ \cdot \\ \cdot \\ \xi_m \end{Bmatrix}. \tag{5}$$

§ 12.5. To find the best estimate s of the parameter σ we have to solve the maximum likelihood equation, P being given in (12.1.5),

$$\frac{\partial}{\partial \sigma} \ln P = \frac{\partial}{\partial \sigma}\left(\text{const} - n \ln \sigma - \frac{[p\epsilon\epsilon]}{2\sigma^2}\right) = -\frac{n}{\sigma} + \frac{[p\epsilon\epsilon]}{\sigma^3} = 0,$$

i.e.,

$$\sigma \approx s_1 = \sqrt{\frac{[pvv]}{n}}. \tag{1}$$

For the same reasons as those discussed in § 11.6–§ 11.7 it is also convenient to use here another estimate, s, having the same relative dispersion as s_1 and being also an asymptotic efficient estimate of σ:

$$\sigma \approx s = \sqrt{\frac{n}{n-m}}\, s_1 = \sqrt{\frac{[pvv]}{n-m}}, \tag{2}$$

where $n - m$ is the degree of freedom since the n v's are subjected to the m constraints (12.3.8).

*§ 12.6. The joint distribution of $\bar{x}_1, \cdots, \bar{x}_m$ and $q = \sqrt{[pvv]}$ is given by the generalization of (7.7.11)

$$d\Phi = \varphi(\bar{x}_1, \cdots, \bar{x}_m, q)\, d\bar{x}_1 \cdots d\bar{x}_m\, dq =$$

$$\left\{\frac{\sqrt{|\boldsymbol{B}|}}{(\sqrt{2\pi}\,\sigma)^m} \exp\left[-\frac{1}{2\sigma^2}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{\xi}})^* \cdot \boldsymbol{B} \cdot (\bar{\boldsymbol{X}} - \bar{\boldsymbol{\xi}})\right] d\bar{x}_1 \cdots d\bar{x}_m\right\} \cdot$$

$$\left\{\frac{1}{[(f-2)/2]!\, 2^{(f-2)/2}} \left(\frac{q}{\sigma}\right)^{f-1} \exp\left[-\frac{q^2}{2\sigma^2}\right] \frac{dq}{\sigma}\right\}, \qquad f = n - m. \quad (1)$$

Exercise. Verify this. (The proof is analogous to that of (7.7.11). Introduce into (12.1.5), instead of $l_1, \cdots, l_n$, the n new variables $\bar{x}_1, \cdots, \bar{x}_m$, q and u_i defined by $l_i = \bar{l}_i - \dfrac{q}{\sqrt{p_i}}\, u_i$, $i = 1, 2, \cdots, n$; use (12.4.4).)

Equation (1) shows that $\bar{x}_1, \cdots, \bar{x}_m$ are independent of q; that $\bar{x}_1, \cdots, \bar{x}_m$ and q are a set of joint sufficient estimates of $\xi_1, \cdots, \xi_m$ and σ (cf. §10.11); and that $\bar{x}_1, \cdots, \bar{x}_m$ are normally distributed with the mean values $\bar{\boldsymbol{\xi}}$ and the moment matrix

$$\boldsymbol{M}^{(\bar{x})} = \mathfrak{M}\{(\bar{x}_r - \xi_r)(\bar{x}_s - \xi_s)\} = \sigma^2 \boldsymbol{B}^{-1} \quad (2)$$

(cf. §7.6), i.e., that the necessary and sufficient condition for $\bar{x}_1, \cdots, \bar{x}_m$ being mutually independent is that $\boldsymbol{B}$ be a diagonal matrix. Finally (1) shows that q has the q-distribution (7.7.12) with the degree of freedom $f = n - m$ in agreement with the fact that the n v's are subject to the m constraints (12.3.8). As shown in §7.7 we thus have, since $s = q/\sqrt{f}$,

$$\mathfrak{M}\{s^2\} = \sigma^2, \qquad \sigma^2\{s^2\} = \frac{2\sigma^4}{n - m} \quad (3)$$

(cf. problem 71), and

$$\mathfrak{M}\{s\} \sim \sqrt{1 - \frac{1}{2(n-m)}}\,\sigma, \quad \sigma^2\{s\} \sim \frac{\sigma^2}{2(n-m)} \quad \text{for}$$

$$f = n - m \gg 1. \quad (4)$$

§12.7. The estimates $\bar{x}_1, \cdots, \bar{x}_m$ and s are, of course, random variables, i.e., subject to statistical fluctuations since n new measurements, $l_1', \cdots, l_n'$, would give somewhat different estimates, $\bar{x}_1', \cdots, \bar{x}_m', s'$. We must, therefore, also estimate the magnitude of these statistical fluctuations. This is done by forming estimates of their dispersions, which we obtain from the moment matrix of $\bar{x}_1, \cdots, \bar{x}_m$, given in (12.6.2).

Thus

$$\sigma\{\bar{x}_j\} = \sqrt{(\boldsymbol{B}^{-1})_{jj}}\,\sigma \approx \sqrt{(\boldsymbol{B}^{-1})_{jj}}\sqrt{\frac{[pvv]}{n-m}}, \qquad j = 1, 2, \cdots, m. \quad (1)$$

We stress that, in general, the $\bar{x}$'s are not uncorrelated since $\boldsymbol{B}^{-1}$ is usually not a diagonal matrix. However, if $\boldsymbol{B}^{-1}$ is a diagonal matrix and, therefore, $\rho\{\bar{x}_i, \bar{x}_j\} = 0$ for $i \neq j$, i.e., the $\bar{x}$'s uncorrelated, $\boldsymbol{B}$ itself is also a diagonal matrix. As shown by (12.6.1) the $\bar{x}$'s are then, furthermore, independent.

Exercise 1. Show that (12.6.2) also follows from (6.4.23), since from (12.1.5)

$$\boldsymbol{M}^{(L)} = \sigma^2 \boldsymbol{P}^{-1} \tag{2}$$

and $\bar{\boldsymbol{X}}$ is a linear function of $\boldsymbol{L}$, given in (12.3.17).

Exercise 2. Show that (12.6.2) also follows from (10.10.15).

Example. In the theory of adjustment it is customary to call a set of linear functions $y_1, \cdots, y_k$ of the direct observations $l_1, \cdots, l_n$ **free functions** if the y's are two-by-two uncorrelated, that is, $\rho\{y_i, y_j\} = 0$ for $i \neq j$. From (6.4.23) and (2) we have

$$\underset{kk}{\boldsymbol{M}^{(Y)}} = \sigma^2 \boldsymbol{F} \cdot \boldsymbol{P}^{-1} \cdot \boldsymbol{F}^*, \tag{3}$$

which shows that the necessary and sufficient condition for the y's being free functions is that $\boldsymbol{F} \cdot \boldsymbol{P}^{-1} \cdot \boldsymbol{F}^*$ is a diagonal matrix. However, since the l's are assumed normally distributed, so are the y's, i.e., (3) shows, furthermore, that the three concepts—mutually independent, two-by-two uncorrelated, and free functions—are equivalent for normally distributed variables (cf. Example 2, § 7.6).

§ 12.8. If we have a set of functions of $\bar{\boldsymbol{X}}$, such as, e.g., $\bar{\boldsymbol{L}}$ given in (12.3.2), and we want to calculate the dispersions of the functions from those of $\bar{\boldsymbol{X}}$ given in (12.6.2), we must remember that the simple variance law (6.4.4) cannot, in general, be applied but that the general variance law (6.4.3) should be used since the $\bar{x}$'s are usually not uncorrelated. However, using the matrix symbolism this is done automatically. Thus if

$$\underset{k1}{\boldsymbol{Z}} = \underset{k1}{\boldsymbol{C}_0} + \underset{km}{\boldsymbol{G}} \cdot \underset{m1}{\bar{\boldsymbol{X}}} \tag{1}$$

is a set of k new random variables, also normally distributed, we have immediately from (6.4.23) and (12.6.2)

$$\underset{kk}{\boldsymbol{M}^{(Z)}} = \sigma^2 \boldsymbol{G} \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{G}^*. \tag{2}$$

Exercise. Show that the same result is obtained if we first express $\boldsymbol{Z}$ as a function of $\boldsymbol{L}$ by means of (12.3.17) and then use (6.4.23) and (12.7.2).

In particular, since from (12.3.2) $\boldsymbol{C}^{(L)} = \boldsymbol{A}$, (2) gives

$$\underset{nn}{\boldsymbol{M}^{(L)}} = \sigma^2 \boldsymbol{A} \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{A}^*, \tag{3}$$

which in general will not be a diagonal matrix. Thus the estimates $\bar{L}$ are usually *not* uncorrelated (i.e., independent). From (3) and (12.5.2)

$$\sigma\{\bar{l}_i\} = \sqrt{(A \cdot \overline{B^{-1} \cdot A^*})_{ii}}\,\sigma \approx \sqrt{(A \cdot \overline{B^{-1} \cdot A^*})_{ii}}\sqrt{\frac{[pvv]}{n-m}},$$

$$i = 1, 2, \cdots, n. \quad (4)$$

As a control we may use the relation

$$\sum_{i=1}^{n} p_i \sigma^2\{\bar{l}_i\} = m\sigma^2, \quad (5)$$

as proved in problem 70.

Example. An easily committed error is that of putting $\sigma\{\bar{l}_i\} = \sigma\{l_i\} = \sigma/\sqrt{p_i}$. That this is erroneous may be seen from the fact that, from (12.1.4) $\sum_{i=1}^{n} p_i \sigma^2\{l_i\} = n\sigma^2$, which does not agree with (5).

§12.9. As for the dispersion of s it follows from (12.6.4) that

$$\sigma\{s\} \sim \frac{\sigma}{\sqrt{2(n-m)}} \approx \frac{s}{\sqrt{2(n-m)}}, \quad (1)$$

if $f = n - m$ is not too small; otherwise we have to use the exact formula (7.7.19).

Exercise 1. Show that from (10.10.14) the asymptotic dispersion of the maximum likelihood estimate s_1, given in (12.5.1), is given by

$$\sigma^2\{s_1\} = \frac{\sigma^2}{2n}, \quad (2)$$

from which (1) also follows.

The final result of the n measurements $l_1, \cdots, l_n$ is now given as:

Estimate of the true value:

$$\xi_j \approx \bar{x}_j \qquad \text{with dispersion:} \quad \sigma\{\bar{x}_j\} \approx \sqrt{(B^{-1})_{jj}}\,s,$$

$$j = 1, 2, \cdots, m. \quad (3)$$

Estimate of the true value:

$$\lambda_i \approx \bar{l}_i \qquad \text{with dispersion:} \quad \sigma\{\bar{l}_i\} \approx \sqrt{(A \cdot B^{-1} \cdot A^*)_{ii}}\,s,$$

$$i = 1, 2, \cdots, n. \quad (4)$$

Estimate of the dispersion:

$$\sigma \approx s \qquad \text{with dispersion:} \quad \sigma\{s\} \approx \frac{s}{\sqrt{2(n-m)}}. \tag{5}$$

(The dispersions of the estimates are often called their **standard** or **mean errors.**)

Exercise 2. Show that

$$[pvv] = \boldsymbol{N}^* \cdot \boldsymbol{P} \cdot \boldsymbol{N} - \boldsymbol{N}^* \cdot \boldsymbol{P} \cdot \boldsymbol{A} \cdot \bar{\boldsymbol{X}} = [pnn] - [pna_1]\bar{x}_1 - \cdots - [pna_m]\bar{x}_m \tag{6}$$

$$[pvv] = \boldsymbol{N}^* \cdot \boldsymbol{P} \cdot \boldsymbol{N} - \bar{\boldsymbol{X}}^* \cdot \boldsymbol{B} \cdot \bar{\boldsymbol{X}} \tag{7}$$

$$[pvv] = [pll] - [p\bar{l}\bar{l}] + 2\boldsymbol{A}_0^* \cdot \boldsymbol{P} \cdot \boldsymbol{V}. \tag{8}$$

These expressions may be used either for the computation of $[pvv]$ or for the control of such a computation.

***Exercise 3.** Show that

$$\frac{x_j - \xi_j}{\sqrt{(\boldsymbol{B}^{-1})_{jj}\boldsymbol{s}}}, \quad j = 1, 2, \cdots, m \quad \text{and} \quad \frac{\bar{l}_i - \lambda_i}{\sqrt{(\boldsymbol{A} \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{A}^*)_{ii}\boldsymbol{s}}}, \quad i = 1, 2, \cdots, n \tag{9}$$

are distributed as a t with $f = n - m$ (cf. §7.9). Find the expressions for the confidence intervals of the parameters $\xi_1, \cdots, \xi_m, \lambda_1, \cdots, \lambda_n$ (cf. §11.9).

Example 1. Let $l_1, \cdots, l_n$ be n, not necessarily equally accurate, measurements of one quantity x; i.e., the number of elements is $m = 1$ and the fundamental equations are

$$\lambda_1 = \lambda_2 = \cdots = \lambda_n = \xi.$$

In this case we have

$$\boldsymbol{A}_0 = \boldsymbol{0}, \qquad \underset{nm}{\boldsymbol{A}} = \underset{n1}{\boldsymbol{A}} = \begin{Bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{Bmatrix}$$

$$\underset{mm}{\boldsymbol{B}} = \underset{11}{\boldsymbol{B}} = \{11 \cdots 1\} \begin{Bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & p_n \end{Bmatrix} \begin{Bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{Bmatrix} = [p],$$

$$\boldsymbol{B}^{-1} = \frac{1}{[p]},$$

$$\underset{mn}{\boldsymbol{C}} = \underset{1n}{\boldsymbol{C}} = \frac{1}{[p]}\{11\cdots 1\}\begin{Vmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & p_n \end{Vmatrix} =$$

$$\left\{\frac{p_1}{[p]}, \frac{p_2}{[p]}, \ldots, \frac{p_n}{[p]}\right\}.$$

Thus we get

$$\xi \approx \bar{x} = \frac{[pl]}{[p]}, \tag{10}$$

which is just the *weighted arithmetic average.* Next

$$\sigma \approx s = \sqrt{\frac{[pvv]}{n-1}} \tag{11}$$

$$\sigma\{\bar{x}\} \approx \sqrt{\frac{[pvv]}{[p](n-1)}} \tag{12}$$

$$\sigma\{s\} \approx \frac{s}{\sqrt{2(n-1)}}. \tag{13}$$

We note that (12) agrees with the result of Exercise 2, § 6.4, and furthermore that it shows that the *weight of the average is also in the general case equal to the sum of the weights of the separate measurements* (cf. § 12.1). For $p_1 = p_2 = \cdots = p_n = 1$, (10)–(13) reduce to the corresponding formulae (11.7.3) and (11.7.4).

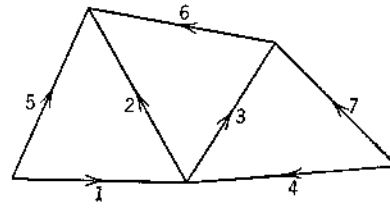

FIG. 21.

Example 2. In a field we measure the altitude differences between 5 points as shown in Fig. 21, where the arrows indicate the directions of the decreases in altitude. The measured values are, in meters,

$$\boldsymbol{L} = \begin{Vmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \\ l_6 \\ l_7 \end{Vmatrix} = \begin{Vmatrix} 20.21 \\ 40.07 \\ 34.17 \\ 35.84 \\ 60.40 \\ 5.87 \\ 69.99 \end{Vmatrix}.$$

As for the weights a closer investigation of the measuring method used in such a surveying problem shows that it is natural to choose them inversely proportional to the corresponding distances. Assuming these to be proportional to 1000, 1110, 910, 1250, 1000, 1110, and 1000, we can take as the weight matrix

$$\underset{77}{\boldsymbol{P}} = \begin{Vmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{Vmatrix}.$$

Choosing as elements the altitude differences 1, 2, 3, and 4, the fundamental equations become

$$\begin{Vmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \\ \lambda_7 \end{Vmatrix} = \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{Vmatrix} \cdot \begin{Vmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{Vmatrix}.$$

Next we find

$$\boldsymbol{B} = \begin{Vmatrix} 2 & 1 & 0 & 0 \\ 1 & 2.8 & -0.9 & 0 \\ 0 & -0.9 & 3 & 1 \\ 0 & 0 & 1 & 1.8 \end{Vmatrix}, \qquad |\boldsymbol{B}| = 17.324$$

$$\boldsymbol{B}^{-1} = \frac{1}{17.324} \begin{Vmatrix} 10.862 & -4.4 & -1.62 & 0.9 \\ -4.4 & 8.8 & 3.24 & -1.8 \\ -1.62 & 3.24 & 8.28 & -4.6 \\ 0.9 & -1.8 & -4.6 & 12.18 \end{Vmatrix}$$

$$\boldsymbol{A}^* \cdot \boldsymbol{P} \cdot \boldsymbol{N} = \begin{Vmatrix} 80.610 \\ 101.746 \\ 102.294 \\ 98.662 \end{Vmatrix},$$

and thus as results

$$\boldsymbol{\Xi} \approx \bar{\boldsymbol{X}} = \begin{Vmatrix} 20.260 \\ 40.090 \\ 34.185 \\ 35.821 \end{Vmatrix} \text{ with dispersions } \begin{Vmatrix} 0.039 \\ 0.035 \\ 0.034 \\ 0.041 \end{Vmatrix}$$

$$\mathbf{\Lambda} \approx \bar{\boldsymbol{L}} = \begin{Bmatrix} 20.260 \\ 40.090 \\ 34.185 \\ 35.821 \\ 60.350 \\ 5.905 \\ 70.006 \end{Bmatrix} \text{ with dispersions } \begin{Bmatrix} 0.039 \\ 0.035 \\ 0.034 \\ 0.041 \\ 0.039 \\ 0.039 \\ 0.040 \end{Bmatrix}$$

and

$$\sigma \approx s = 0.0493 \text{ with dispersion } 0.0201$$

(check all the numerical calculations).

***Example 3.** In Example 1 we have $f = n - m = 7 - 4 = 3$ and $t\,(5\%, 3) = 3.182$ from Table III. Thus the 5% confidence limits are obtained by multiplying the dispersions mentioned by 3.182.

Exercise 4. In practical computations we would in Example 2 choose as elements the true corrections rather than the true values of the altitude differences numbers 1–4 in order to get smaller numbers in the computations. Carry through all the corresponding calculations.

***§ 12.10.** If we want to test whether a measurement is encumbered with a coarse error we may calculate the corresponding relative deviation and by means of the r-distribution test whether it lies within reasonable limits (cf. § 7.10).

Exercise 1. Show that the moment matrix of $\boldsymbol{V}$ is given by

$$\underset{nn}{\boldsymbol{M}^{(V)}} = \sigma^2(\boldsymbol{A} \cdot \boldsymbol{C} - \boldsymbol{E}) \cdot \boldsymbol{P}^{-1} \cdot (\boldsymbol{A} \cdot \boldsymbol{C} - \boldsymbol{E})^* = \sigma^2 \boldsymbol{T}, \tag{1}$$

$\boldsymbol{T}$ being an abbreviation for the complicated matrix given in (1).

The relative deviations we define by

$$r_i = \frac{\bar{l}_i - l_i}{\sqrt{t_{ii}}\, s} = \sqrt{\frac{n-m}{t_{ii}}}\, \frac{v_i}{\sqrt{[pvv]}}, \qquad i = 1, 2, \cdots, n, \tag{2}$$

and it may be shown[1] that the marginal distribution of each r_i is given by the r-distribution (7.10.2) with $f = n - m - 1$ in agreement with the fact that the r's are subjected to $m + 1$ constraints.

Exercise 2. Show that (12.3.8) gives rise to m constraints on the r's. Next show that

$$[ptr^2] = \sum_{i=1}^{n} p_i t_{ii} r_i^2 = n - m. \tag{3}$$

Furthermore the relative deviations may also be used here to check the hypothesis that the observations are normally distributed since

[1] See Arley, *Danske Vid. Selsk., Mat.-fys. Medd.*, XVIII, No. 3, 1940, Chapter II.

we often have several small samples with different true values, weights, and dispersions but with the same degree of freedom f (cf. § 11.21).

ADJUSTMENT BY CORRELATES

§ 12.11. Introducing the matrices

$$\underset{n1}{\boldsymbol{L}} = \begin{Vmatrix} l_1 \\ l_2 \\ \cdot \\ \cdot \\ \cdot \\ l_n \end{Vmatrix}, \quad \underset{n1}{\boldsymbol{V}} = \begin{Vmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{Vmatrix}, \quad \underset{r1}{\boldsymbol{A}_0} = \begin{Vmatrix} a_{10} \\ a_{20} \\ \cdot \\ \cdot \\ \cdot \\ a_{r0} \end{Vmatrix},$$

$$\underset{rn}{\boldsymbol{A}} = \begin{Vmatrix} a_{11} & a_{12} & \cdot\ \cdot\ \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot\ \cdot\ \cdot & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{r1} & a_{r2} & \cdot\ \cdot\ \cdot & a_{rn} \end{Vmatrix}, \tag{1}$$

the equations of condition (12.2.6) may be written

$$\boldsymbol{A}_0 + \boldsymbol{A} \cdot (\boldsymbol{L} + \boldsymbol{V}) = \boldsymbol{0}. \tag{2}$$

As an abbreviation it is customary to put

$$\underset{r1}{\boldsymbol{A}_0} + \boldsymbol{A} \cdot \boldsymbol{L} = \boldsymbol{N}. \tag{3}$$

Thus (2) assumes the form, being analogous to (12.3.4),

$$\boldsymbol{N} + \boldsymbol{A} \cdot \boldsymbol{V} = \boldsymbol{0}. \tag{4}$$

Here we have r equations with $n > r$ unknown v's. Since these v's are not free variables, being subjected to the r constraints (4), the necessary condition that $[pvv]$ = minimum is not in this case that the partial derivatives of $[pvv]$ with respect to all the v's be zero (which would, by the way, lead to $\boldsymbol{V} = \boldsymbol{0}$). The condition is that the total differential of $[pvv]$ be 0, i.e., using (12.3.11),

$$d[pvv] = [2pv\, dv] = 2\boldsymbol{V}^* \cdot \boldsymbol{P} \cdot \boldsymbol{dV} = \boldsymbol{0}, \tag{5}$$

where we have introduced the matrix

$$\underset{n1}{\boldsymbol{dV}} = \begin{Vmatrix} dv_1 \\ dv_2 \\ \cdot \\ \cdot \\ \cdot \\ dv_n \end{Vmatrix}. \tag{6}$$

Differentiating (4) we get

$$\boldsymbol{A} \cdot d\boldsymbol{V} = \boldsymbol{0}. \tag{7}$$

We now eliminate $d\boldsymbol{V}$ between (5) and (7) by means of the method of Lagrangian multipliers called here the **correlates** $k_1, k_2, \cdots, k_r$, which form the *correlate matrix*

$$\underset{r1}{\boldsymbol{K}} = \begin{Bmatrix} k_1 \\ k_2 \\ \cdot \\ \cdot \\ \cdot \\ k_r \end{Bmatrix}. \tag{8}$$

For arbitrary $\boldsymbol{K}$ we have from (5) and (7)

$$\boldsymbol{V}^* \cdot \boldsymbol{P} \cdot d\boldsymbol{V} - \boldsymbol{K}^* \cdot \boldsymbol{A} \cdot d\boldsymbol{V} = (\boldsymbol{V}^* \cdot \boldsymbol{P} - \boldsymbol{K}^* \cdot \boldsymbol{A}) \cdot d\boldsymbol{V} = 0. \tag{9}$$

The method introduced by Lagrange now consists of trying to fix the, as yet, arbitrary correlates so that

$$\boldsymbol{V}^* \cdot \boldsymbol{P} - \boldsymbol{K}^* \cdot \boldsymbol{A} = \boldsymbol{0}, \tag{10}$$

i.e.,

$$\boldsymbol{V} = \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{K}. \tag{11}$$

The n equations (11) with the $n + r$ unknowns, $\boldsymbol{V}$ and $\boldsymbol{K}$, are called the **correlate equations.** Together with the r equations (4) we now have as many equations as unknowns.

Example. The n equations (10) may be arrived at as follows. Since we have r constraints on the n v's we may choose $n - r$ of these as free variables and then express the other remaining r v's as functions of the $n - r$ free ones. Introducing this in $[pvv]$ and taking the partial derivatives with respect to the $n - r$ free v's would then give us the $n - r$ equations which together with the r equations (4) would give us n equations for the unknown v's. These calculations are avoided by the method of Lagrange. Let us assume that as the free v's we have chosen $v_1, \cdots, v_{n-r}$. We then first determine the r correlates so that all the coefficients of the last r differentials in (9) are zero, and next, since $dv_1, \cdots, dv_{n-r}$ are now free, (9) demands for these values of the correlates that the coefficients of the $n - r$ first differentials also be zero. However, we then have in all just the n equations (10).

Introducing (11) into (4), $\boldsymbol{V}$ is eliminated and we get the r equations

$$\boldsymbol{N} + \boldsymbol{A} \cdot \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{K} = \boldsymbol{0}. \tag{12}$$

As an abbreviation we put, analogously to (12.3.13),

$$\underset{rn}{\boldsymbol{A}} \cdot \underset{nn}{\boldsymbol{P}^{-1}} \cdot \underset{nr}{\boldsymbol{A}^*} = \underset{rr}{\boldsymbol{B}}, \tag{13}$$

which is *symmetric* because

$$\boldsymbol{B}^* = (\boldsymbol{A} \cdot \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^*)^* = \boldsymbol{A}^{**} \cdot \boldsymbol{P}^{-1*} \cdot \boldsymbol{A}^* = \boldsymbol{A} \cdot \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* = \boldsymbol{B}. \tag{14}$$

Thus we finally get

$$\boldsymbol{B} \cdot \boldsymbol{K} = -\boldsymbol{N}. \tag{15}$$

These r equations for the r unknown k's are analogous to (12.3.15) and are called the **normal equations** of the system.

Exercise 1. Introducing the r matrices analogous to (12.3.9)

$$\underset{n1}{\boldsymbol{A}_1} = \begin{pmatrix} a_{11} \\ a_{12} \\ \cdot \\ \cdot \\ \cdot \\ a_{1n} \end{pmatrix}, \qquad \underset{n1}{\boldsymbol{A}_2} = \begin{pmatrix} a_{21} \\ a_{22} \\ \cdot \\ \cdot \\ \cdot \\ a_{2n} \end{pmatrix}, \quad \cdots, \quad \underset{n1}{\boldsymbol{A}_r} = \begin{pmatrix} a_{r1} \\ a_{r2} \\ \cdot \\ \cdot \\ \cdot \\ a_{rn} \end{pmatrix}, \tag{16}$$

show that (15) can be written in the form

$$\left[\frac{a_j a_1}{p}\right] k_1 + \left[\frac{a_j a_2}{p}\right] k_2 + \cdots + \left[\frac{a_j a_r}{p}\right] k_r + n_j = 0, \qquad j = 1, 2, \cdots, r, \tag{17}$$

which is the usual form of the normal equations as introduced by Gauss. Sometimes the symbol (yz) is introduced here for $[yz/p]$ and $(a_j n)$ for n_j so that the normal equations become formally identical with those of the adjustment by elements (12.3.16).

From (15) we see that the method of least squares combined with the method of Lagrangian multipliers leads to a unique solution obtained from (15) by multiplication from the left with $\boldsymbol{B}^{-1}$:[1]

$$\boldsymbol{K} = -\boldsymbol{B}^{-1} \cdot \boldsymbol{N}. \tag{18}$$

Finally from (11) we get $\boldsymbol{V}$

$$\boldsymbol{V} = -\boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{N}, \tag{19}$$

and thus

$$\bar{\boldsymbol{L}} = \boldsymbol{L} + \boldsymbol{V} = \boldsymbol{L} - \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{N}. \tag{20}$$

Exercise 2. Show that (20) may also be written

$$\bar{\boldsymbol{L}} = -\boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{A}_0 + (\boldsymbol{E} - \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{A}) \cdot \boldsymbol{L}. \tag{21}$$

[1] Also here the determinant $|\boldsymbol{B}| \neq 0$ for the same reason as in adjustment by elements (cf. p. 187).

§12.12. As discussed in §12.2 adjustment by correlates leads to exactly the same results as adjustment by elements. Consequently we need not prove anew that the v's found in (12.11.19) actually minimize $[pvv]$. Since the number of elements is $m = n - r$ we also have from the previous formulae that

$$\sigma \approx s = \sqrt{\frac{[pvv]}{r}} \tag{1}$$

with the asymptotic dispersion

$$\sigma\{s\} \approx \frac{s}{\sqrt{2r}}. \tag{2}$$

Exercise 1. Show that

$$[pvv] = -\boldsymbol{N}^* \cdot \boldsymbol{K} \tag{3}$$

$$[pvv] = \boldsymbol{K}^* \cdot \boldsymbol{B} \cdot \boldsymbol{K} \tag{4}$$

$$[pvv] = \boldsymbol{N}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{N}, \tag{5}$$

which expressions may be used either for the computation of $[pvv]$ or for the control of such a computation.

Exercise 2. Show that the moment matrices of $\boldsymbol{K}$ and $\bar{\boldsymbol{L}}$ are given by

$$\underset{rr}{\boldsymbol{M}^{(K)}} = \sigma^2 \boldsymbol{B}^{-1} \tag{6}$$

and

$$\underset{nn}{\boldsymbol{M}^{(\bar{L})}} = \sigma^2(\boldsymbol{P}^{-1} - \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{A} \cdot \boldsymbol{P}^{-1}). \tag{7}$$

From (7) we get

$$\sigma\{\bar{l}_i\} = \sqrt{(\boldsymbol{P}^{-1} - \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{A} \cdot \boldsymbol{P}^{-1})_{ii}}\,\sigma \approx \sqrt{(\boldsymbol{P}^{-1} - \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{A} \cdot \boldsymbol{P}^{-1})_{ii}}\,\sqrt{\frac{[pvv]}{r}}. \tag{8}$$

As a control we may use (12.8.5), which owing to $m = n - r$ now becomes

$$\sum_{i=1}^{n} p_i \sigma^2\{\bar{l}i\} = (n - r)\sigma^2. \tag{9}$$

Example. In the problem treated in Example 2, §12.9, we have three constraints between the true values of the seven measurements, viz.,

$$\lambda_1 + \lambda_2 - \lambda_5 = 0$$

$$\lambda_2 - \lambda_3 - \lambda_6 = 0$$

$$\lambda_3 + \lambda_4 - \lambda_7 = 0,$$

i.e.,

$$\underset{0}{\boldsymbol{A}} = \mathbf{0} \text{ and } \underset{37}{\boldsymbol{A}} = \begin{Vmatrix} 1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & -1 \end{Vmatrix},$$

whence

$$\underset{31}{\boldsymbol{N}} = \boldsymbol{A} \cdot \boldsymbol{L} = \begin{Vmatrix} -0.12 \\ 0.03 \\ 0.02 \end{Vmatrix}$$

and

$$\underset{33}{\boldsymbol{B}} = \boldsymbol{A} \cdot \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* = \boldsymbol{A} \cdot \begin{Vmatrix} 1 & 0 & 0 \\ 1.1111 & 1.1111 & 0 \\ 0 & -0.90909 & 0.90909 \\ 0 & 0 & 1.25 \\ -1 & 0 & 0 \\ 0 & -1.1111 & 0 \\ 0 & 0 & -1 \end{Vmatrix} =$$

$$\begin{Vmatrix} 3.1111 & 1.1111 & 0 \\ 1.1111 & 3.1313 & -0.90909 \\ 0 & -0.90909 & 3.1591 \end{Vmatrix}$$

for the values of $\boldsymbol{L}$ and $\boldsymbol{P}$ given in Example 2, § 12.9. This gives

$$\boldsymbol{B}^{-1} = \begin{Vmatrix} 0.37302 & -0.14442 & -0.04156 \\ -0.14442 & 0.40439 & 0.11637 \\ -0.04156 & 0.11637 & 0.35003 \end{Vmatrix}$$

and

$$\boldsymbol{K} = -\boldsymbol{B}^{-1} \cdot \boldsymbol{N} = \begin{Vmatrix} 0.04993 \\ -0.03179 \\ -0.01548 \end{Vmatrix}, \qquad \text{i.e.,}$$

$$\boldsymbol{V} = \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{K} = \begin{Vmatrix} 0.050 \\ 0.020 \\ 0.015 \\ -0.019 \\ -0.050 \\ 0.035 \\ 0.016 \end{Vmatrix},$$

whence finally

$$\bar{\boldsymbol{L}} = \begin{Vmatrix} 20.260 \\ 40.090 \\ 34.185 \\ 35.821 \\ 60.350 \\ 5.905 \\ 70.006 \end{Vmatrix}$$

in agreement with the results obtained in Example 2, §12.9. (Check all the numerical calculations.)

§12.13. We emphasize that the theory of adjustment here developed is valid only under the condition that the fundamental equations (12.2.1) or (12.2.2) are actually linear, i.e., of the form (12.2.5) or (12.2.6). However, in practice, we often meet problems in which the fundamental equations are not born linear. The first problem then is to transform the fundamental equations into linear form. In most cases the conditions mentioned in §6.5 are satisfied, i.e., that the errors are so small that all our functions may to a good approximation be replaced by their corresponding tangent planes, or, in other words, all terms except the linear ones may be neglected in their Taylor series expansions.

For adjustment by elements we choose as elements the true corrections and not the true values (cf. Exercise 4, §12.9). To that purpose we first determine approximate values of the best estimates of the elements, $x_1^0, x_2^0, \cdots, x_m^0$. Next we put

$$\bar{x}_j = x_j^0 + \Delta x_j, \qquad j = 1, 2, \cdots, m \tag{1}$$

and assume the Δx's to be so small that their second and higher powers may be neglected. This condition will, of course, be fulfilled for suitably chosen values of $x_1^0, \cdots, x_m^0$. If the Δx's found by the subsequent calculations turn out not to satisfy this condition we must repeat the calculations with better chosen values of $x_1^0, \cdots, x_m^0$. Expanding by means of Taylor's series the functions in the equations of condition and keeping only the linear terms we get the linearized forms

$$l_i + v_i = f_i(x_1^0 + \Delta x_1, \cdots, x_m^0 + \Delta x_m) = f_i(x_1^0, \cdots, x_m^0) + \left(\frac{\partial f_i}{\partial x_1}\right)_0 \Delta x_1 + \cdots + \left(\frac{\partial f_i}{\partial x_m}\right)_0 \Delta x_m + \cdots, \qquad i = 1, 2, \cdots, n, \tag{2}$$

in which the partial derivatives have to be taken for $\Delta x_1 = \cdots = \Delta x_m = 0$. Introducing the notations

$$f_i(x_1^0, \cdots, x_m^0) = a_{i0}, \qquad i = 1, 2, \cdots, n, \tag{3}$$

and

$$\left(\frac{\partial f_i}{\partial x_j}\right)_0 = a_{ij}, \qquad \begin{matrix} i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, m \end{matrix} \tag{4}$$

we obtain the equations (12.3.2) with Δx_j instead of $\bar{x}_j$. From the theory it follows that the Δx's are normally distributed, but as men-

tioned in Exercise 2, § 7.5, the $\bar{x}$'s will also be normally distributed under our conditions.

For adjustment by correlates we expand directly the functions in Taylor's series after the v's and obtain, neglecting higher powers of the v's,

$$f_i(l_1, \cdots, l_n) + \left(\frac{\partial f_i}{\partial l_1}\right)_0 v_1 + \cdots + \left(\frac{\partial f_i}{\partial l_n}\right)_0 v_n + \cdots = 0,$$

$$i = 1, 2, \cdots, n. \quad (5)$$

In order that the v's be so small that their higher powers may be safely neglected we must assume that the measured values are very close to the best estimates of their true values as obtained from the subsequent calculations. This condition is always satisfied for finer measurements, but for coarser measurements it may very well happen that the v's obtained do not satisfy the condition. In such cases we have to repeat the calculations, expanding not from the measured values but from other more suitable values.

Introducing in (5) the notations

$$f_j(l_1, \cdots, l_n) = n_j, \qquad j = 1, 2, \cdots, m, \quad (6)$$

and

$$\left(\frac{\partial f_j}{\partial l_i}\right)_0 = a_{ji}, \qquad \begin{matrix} i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, m \end{matrix}, \quad (7)$$

we obtain the equations (12.11.4).

Example.[1] In a triangle ABC we have measured the sides a and b as well as the angles A and B and found the values $l_1 = a = 52.3$ cm, $l_2 = b = 33.4$ cm, $l_3 = A = 62°.2$, $l_4 = B = 34°.7$. In this case we have one constraint, $m = 1$, between the four measurements, $n = 4$:

$$f(l_1 + v_1, \cdots, l_4 + v_4) = (l_1 + v_1) \sin (l_4 + v_4) -$$
$$(l_2 + v_2) \sin (l_3 + v_3) = 0.$$

(Inserting the numerical values measured we see that $a \sin B - b \sin A = 0.22 \neq 0$, so that the equation of condition is not satisfied.) Since

$$\frac{\partial f}{\partial l_1} = \sin l_4, \quad \frac{\partial f}{\partial l_2} = -\sin l_3, \quad \frac{\partial f}{\partial l_3} = -l_2 \frac{\pi}{180} \cos l_3,$$

$$\frac{\partial f}{\partial l_4} = l_1 \frac{\pi}{180} \cos l_4,$$

[1] For further examples of non-linear problems we refer to the textbooks on the theory of adjustment given in the list of references.

(5) becomes

$$l_1 \sin l_4 - l_2 \sin l_3 + \sin l_4 \cdot v_1 - \sin l_3 \cdot v_2 - \frac{\pi}{180} l_2 \cos l_3 \cdot v_3 + \frac{\pi}{180} l_1 \cos l_4 \cdot v_4 = 0$$

or, inserting the numerical values,

$$0.22 + 0.57v_1 - 0.88v_2 - 0.27v_3 + 0.75v_4 = 0,$$

i.e.,

$$\underset{11}{\boldsymbol{A}_0} = \boldsymbol{N} = 0.22, \qquad \underset{14}{\boldsymbol{A}} = \{0.57 \quad -0.88 \quad -0.27 \quad 0.75\}.$$

Assuming that the angles have been measured with an accuracy twice as great as that of the sides we have

$$\underset{44}{\boldsymbol{P}} = \begin{Vmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix}.$$

The rest of the calculation now gives

$$\boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* = \begin{Vmatrix} 2.28 \\ -3.52 \\ -0.27 \\ 0.75 \end{Vmatrix}, \quad \underset{11}{\boldsymbol{B}} = \boldsymbol{A} \cdot \boldsymbol{P}^{-1} \cdot \boldsymbol{A}^* = 5.03, \quad \boldsymbol{B}^{-1} = \frac{1}{5.03}$$

$$\underset{11}{\boldsymbol{K}} = -0.0437, \qquad \boldsymbol{V} = \begin{Vmatrix} -0.10 \\ 0.15 \\ 0.01 \\ -0.03 \end{Vmatrix}, \quad \text{i.e.,} \quad \begin{Vmatrix} \bar{a} \\ \bar{b} \\ \bar{A} \\ \bar{B} \end{Vmatrix} = \begin{Vmatrix} 52.20 \text{ cm} \\ 33.55 \text{ cm} \\ 62°.21 \\ 34°.67 \end{Vmatrix}.$$

As a control we calculate $52.20 \sin 34°.67 - 33.55 \sin 62°.21 = 0.02 \sim 0$ within the accuracy of calculation. Finally, calculating the dispersions, we get the results

$$\boldsymbol{\Lambda} \approx \bar{\boldsymbol{L}} = \begin{Vmatrix} 52.20 \text{ cm} \\ 33.55 \text{ cm} \\ 62°.21 \\ 34°.67 \end{Vmatrix} \text{ with dispersions } \begin{Vmatrix} 0.16 \text{ cm} \\ 0.12 \text{ cm} \\ 0°.10 \\ 0°.09 \end{Vmatrix}$$

$$\sigma \approx s = 0.10 \text{ with dispersion } 0.07.$$

(Check all the numerical calculations.)

§12.14. We shall now conclude our discussion of the theory of adjustment by making a remark on when it is more convenient to use adjustment by elements and when it is more convenient to use adjustment by correlates. Since the main work in solving a problem

of adjustment consists in the numerical calculations involved in solving the normal equations,[1] that form of adjustment is, as a rule, preferable which leads to the smallest number of equations. If the number of constraints between the true values of the quantities observed is r, adjustment by elements demands the solution of $m = n - r$ equations, the number of elements being m, while adjustment by correlates demands the solution of r equations, the number of correlates being r. Apart from special cases, the adjustment by elements will thus be preferable when $n - r < r$, i.e., $r > n/2$; adjustment by correlates when $r < n/2$.

§ 12.15. Regression analysis. A most important problem, met in several fields of science, is that of finding by statistical methods the connection between two or more variables. This problem may be treated by various methods, depending on the nature of the problem and the purpose of the analysis. One of these methods is the regression analysis, a special application of the theory of adjustment.

Let us here consider the case of only two variables, $\boldsymbol{x}$ and $\boldsymbol{y}$. Of these $\boldsymbol{x}$ is regarded as an independent variable, $\boldsymbol{y}$ as a dependent variable. $\boldsymbol{x}$ may be either a statistical variable or a non-statistical variable (which, as a rule, only means that the dispersion of $\boldsymbol{x}$ is negligibly small for the problem in question). As previously discussed the latter case is only a special case of the former, the distribution of $\boldsymbol{x}$ then being the causal distribution (cf. Example 1, § 4.3). We assume that the conditional distribution of $\boldsymbol{y}$ is for each value of x normal (or that the directly given variable, $\boldsymbol{y}'$, has been transformed into a new variable, $\boldsymbol{y}$, so that this condition is fulfilled, cf. § 10.3). Furthermore we assume that the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ (cf. § 5.5) is a known function of x, $f(x)$, which contains a certain number of parameters, $\alpha_1, \cdots, \alpha_m$, assumed to enter linearly in $f(x)$:

$$\eta(x) = \mathfrak{M}\{\boldsymbol{y}|x\} = \int_{-\infty}^{\infty} y\varphi(y|x)\,dy = f(x; \alpha_1, \cdots, \alpha_m) = \alpha_1 f_1(x) + \cdots + \alpha_m f_m(x). \quad (1)$$

Finally we assume that the dispersion of $\boldsymbol{y}$ for fixed x, $\sigma^2\{\boldsymbol{y}|x\}$, is either constant or proportional to a known function of x, $g(x)$:

$$\sigma^2\{\boldsymbol{y}|x\} = \int_{-\infty}^{\infty} (y - \eta(x))^2\varphi(y|x)\,dy = \sigma^2 g^2(x). \quad (2)$$

The functions $f_1(x), \cdots, f_m(x)$ and $g(x)$ are either obtained from theoretical arguments or laid down as hypotheses arrived at from a

[1] As for the technique most suitable for such computations see the books mentioned in the list of references.

consideration of a graph in which corresponding x- and y-values are plotted. The problem then is to deduce from the observed y-values, assumed to be stochastically independent, the best estimates, $a_1, \cdots, a_m$, and s for the parameters $\alpha_1, \cdots, \alpha_m$, and σ. *This problem we can at once solve since it obviously may be regarded as a problem of an adjustment by elements, the elements being the parameters* $\alpha_1, \cdots, \alpha_m$. Let $x_1, \cdots, x_n$ be n given values of $\boldsymbol{x}$ (which need not all be different) and $y_1, \cdots, y_n$ the corresponding observed values of $\boldsymbol{y}$. Then we simply have in the formulae of §12.3–§12.9 to make the following transformation:

$$
\begin{aligned}
\{l_1, \cdots, l_n\} &\longleftrightarrow \{y_1, \cdots, y_n\} \\
\{\lambda_1, \cdots, \lambda_n\} &\longleftrightarrow \{\eta_1, \cdots, \eta_n\} \\
\{\bar{l}_1, \cdots, \bar{l}_n\} &\longleftrightarrow \{\bar{y}_1, \cdots, \bar{y}_n\} \\
\{p_1, \cdots, p_n\} &\longleftrightarrow \left\{\frac{1}{g^2(x_1)}, \cdots, \frac{1}{g^2(x_n)}\right\} \\
\{\xi_1, \cdots, \xi_m\} &\longleftrightarrow \{\alpha_1, \cdots, \alpha_m\} \\
\{\bar{x}_1, \cdots, \bar{x}_m\} &\longleftrightarrow \{a_1, \cdots, a_m\} \\
\{a_{10}, \cdots, a_{n0}\} &\longleftrightarrow \{0, \cdots, 0\} \\
\begin{Vmatrix} a_{11} & \cdots & a_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{n1} & \cdots & a_{nm} \end{Vmatrix} &\longleftrightarrow \begin{Vmatrix} f_1(x_1) & \cdots & f_m(x_1) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ f_1(x_n) & \cdots & f_m(x_n) \end{Vmatrix}.
\end{aligned}
\tag{3}
$$

Example 1. The simplest example of a regression analysis is that in which the regression $\eta(x)$ of $\boldsymbol{y}$ on $\boldsymbol{x}$ is a linear function of x

$$\eta(x) = \alpha + \beta x, \tag{4}$$

i.e., in (1) $\alpha_1 = \alpha$, $\alpha_2 = \beta$, $f_1(x) \equiv 1$, $f_2(x) = x$. If for simplicity we assume $g(x) \equiv 1$ we have that the best estimates $\bar{y}_1, \cdots, \bar{y}_n$ of $\eta_1, \cdots, \eta_n$ are given by the following equations of condition

$$
\begin{aligned}
\bar{y}_1 &= a + bx_1 \\
&\cdots \\
\bar{y}_n &= a + bx_n.
\end{aligned}
\tag{5}
$$

Thus

$$\boldsymbol{A}_0 = \mathbf{0}, \quad \underset{n2}{\boldsymbol{A}} = \begin{Vmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{Vmatrix}, \quad \boldsymbol{P} = \boldsymbol{E}, \tag{6}$$

and

$$\underset{22}{\boldsymbol{B}} = \boldsymbol{A}^* \cdot \boldsymbol{P} \cdot \boldsymbol{A} = \begin{Vmatrix} n & [x] \\ [x] & [x^2] \end{Vmatrix}$$

$$|\boldsymbol{B}| = n[x^2] - [x]^2 = n[x^2] - n^2\bar{x}^2 = n[(x - \bar{x})^2], \qquad \bar{x} = \frac{[x]}{n}$$

$$\underset{22}{\boldsymbol{B}^{-1}} = \frac{1}{n[(x - \bar{x})^2]} \begin{Vmatrix} [x^2] & -[x] \\ -[x] & n \end{Vmatrix}. \tag{7}$$

Consequently

$$\begin{Vmatrix} \alpha \\ \beta \end{Vmatrix} \approx \begin{Vmatrix} a \\ b \end{Vmatrix} = \boldsymbol{B}^{-1} \cdot \boldsymbol{A}^* \cdot \boldsymbol{Y} = \begin{Vmatrix} \dfrac{[x^2][y] - [x][xy]}{n[(x - \bar{x})^2]} \\ \dfrac{n[xy] - [x][y]}{n[(x - \bar{x})^2]} \end{Vmatrix} = \begin{Vmatrix} \dfrac{[x^2]\bar{y} - [xy]\bar{x}}{[(x - \bar{x})^2]} \\ \dfrac{[xy] - n\bar{x}\bar{y}}{[(x - \bar{x})^2]} \end{Vmatrix}, \tag{8}$$

$$\bar{y} = \frac{[y]}{n}.$$

Furthermore

$$\sigma \approx s = \sqrt{\frac{[vv]}{n - 2}}, \quad v_i = \bar{y}_i - y_i, \quad i = 1, 2, \cdots, n \tag{9}$$

and

$$\sigma\{a\} \approx \sqrt{\frac{[x^2]}{n[(x - \bar{x})^2]}}\, s, \quad \sigma\{b\} \approx \frac{s}{\sqrt{[(x - \bar{x})^2]}}, \quad \sigma\{s\} \approx \frac{s}{\sqrt{2n - 4}}. \tag{10}$$

Exercise 1. In practice it may be more convenient to write (4) in the form

$$\eta(x) = \alpha + \beta(x - \bar{x}), \qquad \bar{x} = \frac{[x]}{n}. \tag{4a}$$

The reason for preferring (4a) to (4) is partly that this may simplify the numerical calculations and partly that the estimates a and b then become stochastically independent. Prove this latter fact by deducing the formulae corresponding to (5)–(10).

Example 2. Experiments show that, if water runs out of a small hole in a vessel, the square of the volume run out per second, V, is a

linear function of the height, h, of water in the vessel above the hole, i.e.,

$$V^2 = \alpha + \beta h,$$

in which the letters, as always in physical equations, denote the true values of the quantities. Thus V^2 corresponds to our y and h to x. Strictly speaking both V^2 and h are subjected to experimental errors, but those of h may here be regarded as negligibly small. Furthermore V^2 may safely be regarded as normally distributed, and with a constant dispersion. Thus all the conditions of applying a regression analysis are fulfilled, and from the experimental values given in the first two columns in Table 1 the calculations may be carried out according to the formulae of Example 1. However, to simplify the numerical calculations it is more convenient first to make a rough guess, e.g., from a graph, of the values of α and β, say α_0, β_0, i.e., writing

$$V^2 = \alpha_0 + \alpha' + (\beta_0 + \beta')h \qquad \text{or} \qquad V^2 - (\alpha_0 + \beta_0 h) = \alpha' + \beta' h.$$

We now apply our formulae from Example 1 to $y = V^2 - (\alpha_0 + \beta_0 h)$ and $h = x$. With the numerical values

$$\alpha_0 = -0.4320, \qquad \beta_0 = 0.5035$$

we obtain the values given in columns 3–6 of Table 1. We find from the values of Table 1 and the formulae of Example 1

$$a' = -0.007727 \sim -0.0077$$

$$b' = 0.001826 \sim 0.0018,$$

i.e.,

$$\alpha \approx \alpha_0 + a' = -0.4397$$

$$\beta \approx \beta_0 + b' = 0.5053.$$

From these values we obtain the values for $\bar{y}_i$, v_i, and ${v_i}^2$ given in columns 7–9, and thus

$$\sigma \approx s = \sqrt{\frac{287.25}{10-2}} \times 10^{-3} = 0.005992 \sim 0.0060$$

$$\sigma\{a\} \approx \sqrt{\frac{111.0}{10 \times 6.7}}\, s = 0.007712 \sim 0.0077$$

$$\sigma\{b\} \approx \frac{s}{\sqrt{6.7}} = 0.002315 \sim 0.0023$$

$$\sigma\{s\} \approx \frac{s}{\sqrt{20-4}} = 0.001498 \sim 0.0015.$$

TABLE 1

h_i	V_i^2	$\alpha_0 + \beta_0 h_i$	$y_i = V_i^2 - (\alpha_0 + \beta_0 h_i)$	x_i^2	$x_i y_i$	$\bar{y}_i$	$v_i = \bar{y}_i - y_i$	v_i^2
4.530 cm	1.852 cm^6 sec^{-3}	1.8488	$+3.2 \times 10^{-3}$	20.2	$+14.4 \times 10^{-3}$	$+0.51 \times 10^{-3}$	-2.69×10^{-3}	7.24×10^{-6}
4.241	1.703	1.7034	−0.4	17.6	−1.8	−0.02	+0.38	0.14
3.952	1.553	1.5579	−4.9	16.0	−19.6	−0.54	+4.36	19.01
3.663	1.408	1.4123	−4.3	13.7	−15.9	−1.06	+3.24	10.50
3.374	1.274	1.2668	+7.2	11.5	+24.5	−1.59	−8.79	77.26
3.085	1.112	1.1213	−9.3	9.6	−28.8	−2.11	+7.19	51.70
2.796	0.972	0.9758	−3.8	7.8	−10.6	−2.64	+1.16	1.35
2.507	0.832	0.8308	+1.7	6.2	+4.2	−3.17	−4.87	23.72
2.218	0.688	0.6848	+3.2	4.8	+7.0	−3.70	−6.90	47.61
1.929	0.528	0.5392	−11.2	3.6	−21.3	−4.22	+6.98	48.72
$[x] =$			$[y] = -0.0186$	$[x^2] =$	$[xy] =$			$[vv] =$
32.295			$\bar{y} = -0.00186$	111.0	−0.0479			287.25×10^{-6}
$\bar{x} =$								
3.2295								

Exercise 2. Let x denote the velocity (in miles per hour) of an automobile and y the corresponding braking length (in feet). $\mathfrak{M}\{y|x\}$ depends on x, on the mean reaction time, α, of the driver, and on the efficiency of the brakes of the car. In the time α the car will run the distance αx. Furthermore the distance from the time the brakes are put on is found experimentally to be proportional to the square of the velocity. In all we therefore expect the regression of y on x to be given by

$$\eta(x) = \alpha x + \beta x^2.$$

Next the main contribution to the dispersion of y comes from the dispersion σ of the reaction time of the driver about α. Neglecting the contribution from the term βx^2 compared with the contribution from αx we must expect that

$$\sigma\{y|x\} = \sigma x.$$

Find the best estimates for α, β, and σ for the data given in Table 2.

TABLE 2

x (in miles per hour)	4	8	10	12	14	16	18	20	22	24
y (in feet)	2	16	18	20	36	40	42	56	66	70

For a further discussion of regression analysis as well as its generalization to more than one independent variable (which only means substituting for the single variable x in (1) and (2) several variables $x_1, \cdots, x_k$), we must refer to the literature given in the list of references.

§12.16. The χ^2-test of goodness of fit. Another important problem often met is that of testing the agreement between an observed and a theoretical distribution, i.e., of comparing the empirical numbers n_i with the corresponding theoretical numbers $\nu_i = n\varphi_i$ (cf. §§10.4–10.8). For each fixed value of i, n_i is, according to Laplace's formula (8.2.1), for large values of n approximately normally distributed with the parameters $\mathfrak{M}\{n_i\} = \nu_i$, $\sigma\{n_i\} \sim \sqrt{\nu_i}$; i.e., the variable $y_i^2 = \frac{(n_i - \nu_i)^2}{\nu_i}$ is distributed as χ^2 in §7.8 with $f = 1$. As shown in exercise 3 §7.8, the variable $\chi^2 = \sum_{i=1}^{k} y_i^2$ would, for large n, be approximately χ^2-distributed with $f = k$ if the y_i's were independent variables. In fact they are not, because they satisfy the linear relation $[\sqrt{\nu_i} y_i] = [n_i] - [\nu_i] = n - n = 0$ and possibly other relations such that the n_i's have a definite mean or dispersion. However, it may be shown that nevertheless the quantity

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \nu_i)^2}{\nu_i}, \quad \nu_i = n\varphi_i, \quad f = k - m \tag{1}$$

is, for large values of n, approximately χ^2-distributed, but with a number of degrees of freedom $f = k - m$, i.e., the number, k, of groups into which the observations are classified minus the number, m, of relations between the n_i's.[1] Thus, for large n, the probability $P(\chi^2 \geq \chi_P^2)$ will be given approximately by Table V giving χ_P^2 as a function of f and P. If in one single sample the value of χ^2 exceeds the χ_P^2 corresponding to P, say $P = 5\%$, it is customary to consider the observations to show a significant deviation form the theoretical distribution. (It is here assumed in practice that the classification is such that all $n_i > 5$.) *We wish to stress that such a test, like all other purely statistical tests, must be applied with a certain tact.* First, the distribution of χ^2 defined in (1) is only asymptotically χ^2-distributed. Second, this test does not take into account the signs and order of the deviations $n_i - \nu_i$, and thus the test does not disclose possible systematic deviations. Third, the test is only negative, giving only a possibility for rejecting theoretical distributions showing too strong disagreement with observations. Fourth, it is completely arbitrary which limit, usually 5%, is chosen as the level of significance, just as it is also completely arbitrary that this specific test is chosen among many other possibilities. Applying any such tests mechanically may easily veil the fundamental fact that, as stressed throughout this book (cf., e.g., § 9.3), *it is a subjective question whether or not a mathematical model fits observational facts.* The degree of agreement can only be settled by each person and for each problem separately. (E.g., in spectroscopy we would demand a much higher degree of agreement than in biology.)

Example. In the example in § 10.4 we first have to group together the last four classes, since for each of them $n_i < 5$. Having done this we get $\chi^2 = 12.88$ (check!). Having $k = 12$ classes and $m = 2$ relations between the n_i's, viz., total number $n = 2608$, and mean value $\mu = 3.87$, we have $f = 12 - 2 = 10$. From Table V we then get $\chi_{0.05}^2 = 18.307$, and since the χ^2 observed lies well within this limit the test does not give any reason to reject the theoretical distribution.

Exercise. Show that in the example in § 10.5 we get for the azimuth deviation $\chi^2 = 0.62$ with $f = 5 - 3 = 2$, and for the height deviation $\chi^2 = 1.01$ with $f = 7 - 3 = 4$. What does the test show in this case?

[1] See, e.g., Cramér: *Mathematical methods of statistics*, Chap. 30. Why does the above statement not follow directly from our § 12.6?

APPENDIX 1

$n!$

By $n!$ (read factorial n) we understand the number[1]

$$n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1,$$

if n is a positive integer. By partial integration we find that for such values of n

$$n! = \int_0^\infty t^n e^{-t}\,dt \qquad (e = 2.71828\cdots);$$

putting the integral equal to $f(n)$, we find

$$f(n) = -\int_0^\infty t^n\,d(e^{-t}) = \left|-t^n e^{-t}\right|_0^\infty + n\int_0^\infty t^{n-1}e^{-t}\,dt = nf(n-1),$$

since $\lim_{t\to\infty} t^n e^{-t} = 0$. Repeating this process we find $f(n) = n(n-1)\cdots 3\cdot 2\cdot \int_0^\infty e^{-t}\,dt = n!$. It may be shown that the integral exists for all values of $n > -1$, and for an arbitrary $n > -1$ we therefore define $n!$ as the value of the integral. The function $\Gamma(n) = (n-1)!$ is called the gamma function. Thus in particular $0! = \int_0^\infty e^{-t}\,dt = 1$. (Furthermore it may be shown that $z!$ may be defined for all complex values of z such that the relation $z! = z\cdot(z-1)!$ found above is satisfied.)

It may be shown that for large values of n we have the so-called Stirling's formula

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left[\frac{\theta}{12n}\right], \qquad |\theta| < 1,$$

where θ is a certain number depending on n. If we put $\theta = 0$ the error is already less than 10 % for $n = 1$.

By means of the substitution $ax^2 = t$, $a > 0$, the integral

$$\int_0^\infty x^n e^{-ax^2}\,dx,$$

which is often needed in probability, may be expressed by means of the previous integral, and vice versa. We obtain

[1] For a more detailed discussion as well as proofs see the list of references.

$$\int_0^\infty x^n e^{-ax^2}\,dx = \frac{1}{a^{(n+1)/2}} \int_0^\infty t^{n/2} e^{-t} \frac{dt}{2\sqrt{t}} = \frac{1}{2a^{(n+1)/2}} \int_0^\infty t^{(n-1)/2} e^{-t}\,dt =$$

$$\frac{1}{2a^{(n+1)/2}} \left(\frac{n-}{2} \right.$$

Putting $n = 0$, $a = 1$ and $x^2 = t^2/2$ we thus find from (7.1.2)

$$\tfrac{1}{2}(-\tfrac{1}{2})! = \int_0^\infty e^{-x^2}\,dx = 1/(2\sqrt{2}) \int_{-\infty}^{\infty} e^{-(t^2/2)}\,dt = \sqrt{\pi}/2,$$

i.e., $(-\tfrac{1}{2})! = \sqrt{\pi}$. From this we find $(\tfrac{1}{2})! = \tfrac{1}{2}(-\tfrac{1}{2})! = \sqrt{\pi}/2$, $(\tfrac{3}{2})! = \tfrac{3}{2}(\tfrac{1}{2})! = \tfrac{3}{4}\sqrt{\pi}$, $(\tfrac{5}{2})! = \tfrac{5}{2}(\tfrac{3}{2})! = \tfrac{15}{8}\sqrt{\pi}$, and so forth.

APPENDIX 2

Matrix Theory

We shall here shortly review the fundamental concepts and theorems of the theory of matrices.[1]

1. By a **matrix** we understand a rectangular scheme of numbers, the *elements*, denoted here by small letters, while the matrices themselves are denoted by capital, boldface letters:

$$\boldsymbol{A} = \underset{mn}{\boldsymbol{A}} = \{a_{rs}\} = \begin{Vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{Vmatrix}.$$

We say that $\underset{mn}{\boldsymbol{A}}$ has m *rows* and n *columns*. In particular, for $n = 1$, $\boldsymbol{A}$ is called a *column matrix*, for $m = 1$ a *row matrix*. For $m = n = 1$, $\boldsymbol{A}$ reduces to a number. The elements are also denoted $(A)_{rs}$.

2. We say that two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are identical, $\boldsymbol{A} = \boldsymbol{B}$, if they have the same number of rows, the same number of columns, and $a_{rs} = b_{rs}$ for all r and s.

3. A *zero matrix*, $\boldsymbol{0}$, is a matrix all elements of which are 0, $a_{rs} = 0$. We stress that there exist infinitely many zero matrices, one for each value of m and n.

4. Interchanging rows and columns in $\underset{mn}{\boldsymbol{A}}$ we get a new matrix with n rows and m columns, called the *transposed* matrix of $\boldsymbol{A}$ and denoted $\underset{nm}{\boldsymbol{A}^*}$.

5. By the matrix $\alpha\boldsymbol{A}$, α an arbitrary number, we understand the matrix obtained when all the elements of $\boldsymbol{A}$ are multiplied by α, i.e.,

$$\alpha\boldsymbol{A} = \alpha\{a_{rs}\} = \{\alpha a_{rs}\}.$$

6. If two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of rows and the same number of columns we define their *sum*

$$\underset{mn}{\boldsymbol{A}} + \underset{mn}{\boldsymbol{B}} = \underset{mn}{\boldsymbol{C}}$$

as the matrix having the elements

$$c_{rs} = a_{rs} + b_{rs}.$$

By this definition all the usual rules for sums and differences are valid.

7. If the number of columns of one matrix $\underset{mn}{\boldsymbol{A}}$ is equal to the number of rows of a second matrix $\underset{np}{\boldsymbol{B}}$ we define their *product*

$$\underset{mn}{\boldsymbol{A}} \cdot \underset{np}{\boldsymbol{B}} = \underset{mp}{\boldsymbol{C}}$$

[1] For a more detailed discussion as well as proofs see the list of references.

as the matrix having the elements

$$c_{rs} = a_{r1}b_{1s} + a_{r2}b_{2s} + \cdots + a_{rn}b_{ns} = \sum_{i=1}^{n} a_{ri}b_{is}.$$

Thus a product has as many *rows* as the *first* factor and as many *columns* as the *second* factor. For this so-called *row-column-product* the *associative law* is valid

$$\underset{mn}{(\boldsymbol{A}} \cdot \underset{np}{\boldsymbol{B})} \cdot \underset{pq}{\boldsymbol{C}} = \underset{mn}{\boldsymbol{A}} \cdot (\underset{np}{\boldsymbol{B}} \cdot \underset{pq}{\boldsymbol{C}})$$

as well as the *distributive law*

$$\underset{mn}{\boldsymbol{A}} \cdot (\underset{np}{\boldsymbol{B}} + \underset{np}{\boldsymbol{C}}) = \underset{mn}{\boldsymbol{A}} \cdot \underset{np}{\boldsymbol{B}} + \underset{mn}{\boldsymbol{A}} \cdot \underset{np}{\boldsymbol{C}}$$

$$(\underset{mn}{\boldsymbol{A}} + \underset{mn}{\boldsymbol{B}}) \cdot \underset{np}{\boldsymbol{C}} = \underset{mn}{\boldsymbol{A}} \cdot \underset{np}{\boldsymbol{C}} + \underset{mn}{\boldsymbol{B}} \cdot \underset{np}{\boldsymbol{C}}.$$

However, the *commutative law*, $\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{B} \cdot \boldsymbol{A}$, is in general *not* valid. For example, if the number of columns of $\boldsymbol{B}$ differs from the number of rows of $\boldsymbol{A}$, the second product $\boldsymbol{B} \cdot \boldsymbol{A}$ cannot even be formed although the first one can.

8. For a product we have

$$(\underset{mn}{\boldsymbol{A}} \cdot \underset{np}{\boldsymbol{B}})^* = \underset{pn}{\boldsymbol{B}^*} \cdot \underset{nm}{\boldsymbol{A}^*},$$

which rule may immediately be generalized to products containing more than two factors.

9. A *quadratic* matrix, $\underset{nn}{\boldsymbol{A}}$, is a matrix for which the number of rows is equal to the number of columns.

10. A *quadratic* matrix is called symmetric if $\boldsymbol{A}^* = \boldsymbol{A}$, i.e., $a_{rs} = a_{sr}$ for all r and s. For an arbitrary matrix $\boldsymbol{A}$ both products $\boldsymbol{A} \cdot \boldsymbol{A}^*$ and $\boldsymbol{A}^* \cdot \boldsymbol{A}$ exist and are symmetric, although they need not be identical.

11. A *diagonal* matrix is a quadratic matrix for which $a_{rs} = 0$ when $r \neq s$. The elements a_{rr} are called the *diagonal elements*.

12. A *unit matrix* is a diagonal matrix for which $a_{rr} = 1$. It is denoted $\boldsymbol{E}$ (or 1), and we have

$$\boldsymbol{A} \cdot \boldsymbol{E} = \boldsymbol{E} \cdot \boldsymbol{A} = \boldsymbol{A}$$

for an arbitrary matrix $\boldsymbol{A}$ for which the products with $\boldsymbol{E}$ exist. We stress that there exist infinitely many unit matrices, one for each value of n.

13. By the *determinant* $|\boldsymbol{A}|$ of a quadratic matrix $\boldsymbol{A}$ we understand the determinant $|a_{rs}|$. In particular, if $\boldsymbol{A}$ is a diagonal matrix, $|\boldsymbol{A}| = a_{11}a_{22} \cdots a_{nn}$. We always have $|\boldsymbol{A}^*| = |\boldsymbol{A}|$. We also have $|\boldsymbol{A} \cdot \boldsymbol{B}| = |\boldsymbol{A}| \cdot |\boldsymbol{B}|$.

14. If $|\boldsymbol{A}| \neq 0$ there exists one and only one matrix, called the *reciprocal* matrix of $\boldsymbol{A}$ and denoted $\boldsymbol{A}^{-1} = \{(\boldsymbol{A}^{-1})_{rs}\}$, for which

$$\boldsymbol{A} \cdot \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} \cdot \boldsymbol{A} = \boldsymbol{E}.$$

If $n > 1$ the elements of $\boldsymbol{A}^{-1}$ are given by

$$(\boldsymbol{A}^{-1})_{sr} = \frac{K_{rs}}{|\boldsymbol{A}|},$$

where K_{rs} is the cofactor of the element a_{rs} in the determinant $|\boldsymbol{A}|$, i.e., $(-1)^{r+s}$ times the minor obtained from $|\boldsymbol{A}|$ by taking away the rth row and the sth column.

In particular, if A is a diagonal matrix, A^{-1} is also a diagonal matrix, and

$$(A^{-1})_{rr} = \frac{1}{a_{rr}}.$$

15. If $|A| \neq 0$ the *matrix equation*

$$A \cdot X = B$$

has one and only one solution obtained by multiplying from the left with A^{-1}:

$$A^{-1} \cdot A \cdot X = X = A^{-1} \cdot B.$$

In the same way the matrix equation

$$Y \cdot A = B$$

has for $|A| \neq 0$ one and only one solution obtained by multiplying from the right with A^{-1}:

$$Y \cdot A \cdot A^{-1} = Y = B \cdot A^{-1}.$$

We stress that in general $X \neq Y$.

16. If A^{-1} and B^{-1} exist, then

$$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}.$$

17. A homogeneous, quadratic form $\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$, $a_{ij} = a_{ji}$, may be written as $X^* \cdot A \cdot X$, where $A = \{a_{rs}\}$ and X is the column matrix with the elements $x_1, \cdots, x_n$.

18. An *orthogonal* matrix is a quadratic matrix for which

$$A \cdot A^* = A^* \cdot A = E.$$

For an orthogonal matrix $|A| = \pm 1$ and $A^* = A^{-1}$.

19. Introducing in $X^* \cdot A \cdot X$ new variables by letting

$$\underset{n1}{X} = \underset{nn}{F} \cdot \underset{n1}{Y}$$

we get

$$X^* \cdot A \cdot X = Y^* \cdot B \cdot Y, \qquad B = F^* \cdot A \cdot F.$$

For an arbitrary, symmetric matrix there exists an orthogonal matrix F such that B is a diagonal matrix. In such a case we have $|A| = |B| = b_{11} b_{22} \cdots b_{nn}$.

TABLE I

The Normal Distribution (§7.1)

t	$\psi(t)$	$\Psi(t)$	$2\Psi(t) - 1$
0.0	0.39894	0.50000	0.00000
0.1	.39695	.53983	.07966
0.2	.39104	.57926	.15852
0.3	.38139	.61791	.23582
0.4	.36827	.65542	.31084
0.5	0.35207	0.69146	0.38292
0.6	.33322	.72575	.45150
0.7	.31225	.75804	.51608
0.8	.28969	.78814	.57628
0.9	.26609	.81594	.63188
1.0	0.24197	0.84134	0.68268
1.1	.21785	.86433	.72866
1.2	.19419	.88493	.76986
1.3	.17137	.90320	.80640
1.4	.14973	.91924	.83848
1.5	0.12952	0.93319	0.86638
1.6	.11092	.94520	.89040
1.7	.09405	.95543	.91086
1.8	.07895	.96407	.92814
1.9	.06562	.97128	.94256
2.0	0.05399	0.97725	0.95450
2.1	.04398	.98214	.96428
2.2	.03547	.98610	.97220
2.3	.02833	.98928	.97856
2.4	.02239	.99180	.98360
2.5	0.01753	0.99379	0.98758
2.6	.01358	.99534	.99068
2.7	.01042	.99653	.99306
2.8	.00792	.99744	.99488
2.9	.00595	.99813	.99626
3.0	0.00443	0.99865	0.99730
3.1	.00327	.99903	.99806
3.2	.00238	.99931	.99862
3.3	.00172	.99952	.99904
3.4	.00123	.99966	.99932
3.5	0.00087	0.99977	0.99954
3.6	.00061	.99984	.99968
3.7	.00042	.99989	.99978
3.8	.00029	.99993	.99986
3.9	.00020	.99995	.99990
4.0	0.00013	0.99997	0.99994

TABLE II[1]

The Normal Distribution (§7.4)

P	a
1.0	0
0.9	0.12566
0.8	0.25335
0.7	0.38532
0.6	0.52440
0.5	0.67449
0.4	0.84162
0.3	1.03643
0.2	1.28155
0.1	1.64485
0.05	1.95996
0.01	2.57583
0.001	3.29053
10^{-4}	3.89059
10^{-5}	4.41717
10^{-6}	4.89164
10^{-7}	5.32672
10^{-8}	5.73073
10^{-9}	6.10941

[1] This table is reprinted from Table I of Fisher and Yates, *Statistical Tables*, Oliver and Boyd, by kind permission of the authors and publishers.

TABLE III[1]

The t-Distribution (§7.9)

P / f	0.1	0.05	0.01	0.001
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.598
3	2.353	3.182	5.841	12.941
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.859
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.405
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.767
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.690
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.659
30	1.697	2.042	2.750	3.646
35	1.689	2.030	2.724	3.591
40	1.684	2.021	2.704	3.551
45	1.679	2.014	2.689	3.522
50	1.676	2.008	2.677	3.497
60	1.671	2.000	2.660	3.460
70	1.667	1.995	2.648	3.436
80	1.664	1.990	2.639	3.416
90	1.662	1.987	2.632	3.401
100	1.660	1.984	2.626	3.391
120	1.658	1.980	2.617	3.373
∞	1.645	1.960	2.576	3.291

[1] This table is abbreviated from Table III of Fisher and Yates, *Statistical Tables*, Oliver and Boyd, by kind permission of the authors and publishers.

TABLE IV[2]

The r-Distribution (§7.10)

P / f	0.1	0.05	0.01	0.001
1	1.397	1.409	1.414	1.414
2	1.559	1.645	1.715	1.730
3	1.611	1.757	1.918	1.982
4	1.631	1.814	2.051	2.178
5	1.640	1.848	2.142	2.329
6	1.644	1.870	2.208	2.447
7	1.647	1.885	2.256	2.540
8	1.648	1.895	2.294	2.616
9	1.649	1.903	2.324	2.678
10	1.649	1.910	2.348	2.730
11	1.649	1.916	2.368	2.774
12	1.649	1.920	2.385	2.812
13	1.649	1.923	2.399	2.845
14	1.649	1.926	2.412	2.874
15	1.649	1.928	2.423	2.899
16	1.649	1.931	2.432	2.921
17	1.649	1.933	2.440	2.941
18	1.649	1.935	2.447	2.959
19	1.649	1.936	2.454	2.975
20	1.649	1.937	2.460	2.990
21	1.649	1.938	2.465	3.003
22	1.648	1.940	2.470	3.015
23	1.648	1.941	2.475	3.026
24	1.648	1.941	2.479	3.037
25	1.648	1.942	2.483	3.047
26	1.648	1.943	2.487	3.056
27	1.648	1.943	2.490	3.064
28	1.648	1.944	2.492	3.071
29	1.648	1.945	2.495	3.078
30	1.648	1.945	2.498	3.085
35	1.648	1.948	2.509	3.113
40	1.648	1.949	2.518	3.134
45	1.647	1.950	2.524	3.152
50	1.647	1.951	2.529	3.166
60	1.646	1.953	2.537	3.186
70	1.646	1.954	2.542	3.201
80	1.646	1.955	2.547	3.211
90	1.646	1.956	2.550	3.220
100	1.646	1.956	2.553	3.227
120	1.646	1.957	2.556	3.237
∞	1.645	1.960	2.576	3.291

[2] This table is abbreviated from Table 1 in Arley, *Danske Vid. Selsk. Mat.-fys. Medd.*, Vol. XVIII, No. 3, 1940.

TABLE V[1]

The χ^2-Distribution (§ 7.8)

P / f	0.95	0.1	0.05	0.01	0.001
1	0.00393	2.706	3.841	6.635	10.827
2	0.103	4.605	5.991	9.210	13.815
3	0.352	6.251	7.815	11.341	16.268
4	0.711	7.779	9.488	13.277	18.465
5	1.145	9.236	11.070	15.086	20.517
6	1.635	10.645	12.592	16.812	22.457
7	2.167	12.017	14.067	18.475	24.322
8	2.733	13.362	15.507	20.090	26.125
9	3.325	14.684	16.919	21.666	27.877
10	3.940	15.987	18.307	23.209	29.588
11	4.575	17.275	19.675	24.725	31.264
12	5.226	18.549	21.026	26.217	32.909
13	5.892	19.812	22.362	27.688	34.528
14	6.571	21.064	23.685	29.141	36.123
15	7.261	22.307	24.996	30.578	37.697
16	7.962	23.542	26.296	32.000	39.252
17	8.672	24.769	27.587	33.409	40.790
18	9.390	25.989	28.869	34.805	42.312
19	10.117	27.204	30.144	36.191	43.820
20	10.851	28.412	31.410	37.566	45.315
21	11.591	29.615	32.671	38.932	46.797
22	12.338	30.813	33.924	40.289	48.268
23	13.091	32.007	35.172	41.638	49.728
24	13.848	33.196	36.415	42.980	51.179
25	14.611	34.382	37.652	44.314	52.620
26	15.379	35.563	38.885	45.642	54.052
27	16.151	36.741	40.113	46.963	55.476
28	16.928	37.916	41.337	48.278	56.893
29	17.708	39.087	42.557	49.588	58.302
30	18.493	40.256	43.773	50.892	59.703

[1] This table is abbreviated from Table IV of Fisher and Yates, *Statistical Tables*, Oliver and Boyd, by kind permission of the authors and publishers.

TABLE VI[1]

The w^2-(or F-)Distribution (§7.11); $P=0.05$

f_1 / f_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

[1] This table is abbreviated from Table V of Fisher and Yates, *Statistical Tables*, Oliver and Boyd, by kind permission of the authors and publishers.

PROBLEMS

Problems to Chapters 1 through 3

1. From each of 6 decks of cards, each containing 52 cards, 1 card is drawn at random. Find the probability of 4 being red and 2 being black cards.

2. From one deck of 52 cards, 6 are drawn at random. Find the probability of 4 being red and 2 being black cards.

3. Find the probability of getting exactly once the result 6 in 4 throws with a die.

4. The sum of a column of numbers is a number with 7 digits. Find the probability of guessing this number.

5. Four positive integers are chosen at random. Find the probability of their having a common factor. (Use the formula

$$\prod_{i=1}^{\infty}\left(1-\frac{1}{p_i{}^4}\right)=\frac{90}{\pi^4},$$

where p_i denotes the ith prime number.)

6. A deer runs with constant velocity v in the direction away from a hunter. The probability of the hunter hitting the deer when shooting from the distance $d(>v)$ is assumed to be v^2/d^2. He fires one shot when the deer is at a distance $d = 2v$ and, if this is a failure, one at the distance $d = 3v$, and so forth. However, at most he fires n shot. What is the probability of the hunter's shooting down the deer?

7. Deduce the general formulae that are analogous to the addition law IV, for

$$P(A+B+C) \qquad \text{and} \qquad P(A+B+C+D).$$

8. Find the probability that *at least* 1 player in a bridge hand will get 13 cards of the same color. (Use problem 7.)

9. In bridge a player A has 2 aces. Find the probability of his partner having at least 1 of the 2 remaining aces.

10. An event has the probability 1/10 of occurring. Find the probability that the event occurs *at least* 10 times in 100 trials.

11. A shipowner expects 2 boats with bananas. The statistics show that 1% of all banana cargos on boats are spoiled in transit. Find the probability of (*a*) both freights arriving unspoiled, (*b*) of only 1 arriving unspoiled, and (*c*) of neither of them arriving unspoiled.

12. What is the probability of torpedoing a ship when in all 3 torpedoes may be fired and the probability of hitting with 1 torpedo is 0.25?

13. A man stands at the origin of a number axis and throws heads and tails with a coin. If the result is heads, he walks 1 step of unit length to the right and, if it is tails, 1 to the left. If x denotes his abscissa after 10 throws, what are the possible values of x, and what are their probabilities?

14. A torpedo boat is at a distance d from a target and has the probability 0.05 and fires 3 torpedoes. Assuming the probability of hitting to be inversely pro-

portional to the square of the distance, find the probability (*a*) of 3 hits, p_a; (*b*) of 2 hits, p_b; (*c*) of at least 2 hits, p_c; (*d*) of at least 1 hit, p_d; and (*e*) of no hits, p_e.

15. The probability of hitting a fixed target with 1 gun at the distance d is given as 0.5. At what distance from a target must a battery of 4 guns be placed in order that the probability of exactly 2 hits in a volley equals 3/32? (The probability of hitting is assumed to vary inversely proportional with the square of the distance.)

16. A bus stop is passed every 4th minute by a bus of line A and every 6th minute by one of line B. Assuming the departing times of the two lines to be independent, find the probability (*a*) that the first bus arriving is of line A, and (*b*) that either bus will arrive within 2 minutes.

17. Two batteries A and B with the same probability of hitting, p, each fire n shots against a common target. Assuming exactly 2 hits among the $2n$ shots, find the probability (*a*) of both hits coming from battery A, (*b*) of 1 hit coming from A, the other from B, and (*c*) of both hits coming from B.

18. The probability that a man who has just passed his pth birthday dies before his next birthday is P_p. Find the probability that a man who has just passed his 50th birthday will die within the next 5 years.

19. A and B play heads and tails. A throws first, and wins if the result is heads. If it is tails B throws and wins if the result is heads. If it is tails A throws again, and so forth. If after $2n$ throws neither A nor B has won, a third person C has won. Find the probabilities P_A, P_B, P_C, of A, B, and C winning.

20. In a card game only the cards $1, 2, \cdots, 10$ of one definite color are used. The game is played by 5 persons and consists of having the players, in a definite order, get 2 cards each. A player wins or loses depending on whether or not the sum of the numbers of his 2 cards is equal to 11. What is the probability (*a*) of p of the players, designated beforehand, getting 11, the $5 - p$ remaining players not? (*b*) of just the player k being the first to get 11? (*c*) of none of the 5 players getting 11?

21. Let the probability that the weather on one day is of the same kind (rain or no rain) as on the previous day be p, and let P be the probability of rain on the first day of the year. Find the probability, P_n, of rain on the nth day, and find the limit of P_n for $n \to \infty$.

Problems to Chapters 4 through 6

22. Three batteries with the probabilities of hitting 0.1, 0.2, and 0.3 respectively each fire 1 shot. Find the probability of each of the possible number of hits. Find the distribution function of the shots.

23. Let the random variable x denote the number of throws with a coin until the result "heads" appears for the first time. Find the distribution function, the mean value, and the dispersion.

24. Let x have the probability density

$$\varphi(t) = \begin{cases} 0 & \text{for} \quad -\infty < t \leqq -\dfrac{1}{n} \\ n + n^2 t & \text{for} \quad -\dfrac{1}{n} \leqq t \leqq 0 \\ n - n^2 t & \text{for} \quad 0 \leqq t \leqq \dfrac{1}{n} \\ 0 & \text{for} \quad \dfrac{1}{n} \leqq t < \infty. \end{cases}$$

Find $\Phi(t)$, μ, and σ. Draw the graphs of $\varphi(t)$ and $\Phi(t)$. Show that, for $n \to \infty$, $\Phi(t) \to \epsilon(t)$.

25. Two numbers are chosen at random between 0 and 1. Find the probability that their product x will be less than t, $0 < t < 1$. Find μ and σ for the random variable x.

26. On a circle three points A, B, and C are chosen at random. Find the probability of the triangle ABC being acute angled.

27. Three numbers a, b, and c are chosen at random between 0 and 1. Find the probability of the equation

$$ax^2 + 2bx + c = 0$$

having real roots.

28. A random variable x is assumed binomially distributed such that $\mu = 2$ and $\sigma^2 = \frac{3}{4}$. Find the probabilities

$$P(3 < x < 5), \quad P(3 \leqq x < 5), \quad P(3 < x \leqq 5), \quad \text{and} \quad P(3 \leqq x \leqq 5).$$

29. Five enumerated persons are placed at random on 5 chairs enumerated with the same numbers as the persons. Let x be the random variable which denotes the number among the 5 persons who are placed on chairs with a number identical with their own. Find the distribution function of x as well as μ and σ.

30. An urn contains 4 red and 3 black balls. Four balls are drawn at random. Let the random variable x denote the number of red balls among the 4 drawn. Find the mean value and dispersion of x. Solve the same problem when the balls are drawn successively and each ball is put back in the urn before the next drawing.

31. From a collection of 6 balls, enumerated from 1 to 6, 3 are drawn at random. Let the random variable x denote the largest of the numbers so obtained. Draw the graph of x's distribution function, and find its mean value and dispersion.

32. A random variable x has a given probability density $\varphi(t)$. Find the probability density of $y = 1/x$ and $z = \cos x$. In particular, assume $\varphi(t)$ to be normal.

33. A random variable x has the probability density

$$\varphi(t) = \frac{a}{e^t + e^{-t}}.$$

Determine the value of a. Next find the probability of the largest of two observations being less than 1.

34. Let x be a random variable with an arbitrary distribution having finite mean value μ and dispersion σ. Find another random variable x' so that x' has a rectangular distribution with the same μ and σ as x.

35. A random variable x which can assume only positive values has the logarithmico-normal distribution

$$d\Phi = \varphi(t)\, dt = \frac{1}{\sqrt{2\pi}\,\beta} \exp\left[-\frac{(\ln t - \alpha)^2}{2\beta^2}\right]\frac{dt}{t}.$$

Find the mean value and dispersion of x.

36. A random variable x which can assume only positive values has the probability density

$$d\Phi = \varphi(t)\, dt = e^{-t}\, dt.$$

Let $x_1, \cdots, x_\nu$ be ν independent observations of x. Find the distribution of

$$z_\nu = x_1 + x_2 + \cdots + x_\nu$$

and the mean value and dispersion of z_ν.

37. Let x be a random variable with given distribution function $\Phi(t)$ for which $\mu = 0$ and σ is finite. Show that

$$\Phi(t) \begin{cases} \leqq \dfrac{\sigma^2}{\sigma^2 + t^2} & \text{for} \quad t < 0 \\ \geqq \dfrac{t^2}{\sigma^2 + t^2} & \text{for} \quad t > 0 \end{cases}.$$

Furthermore, show by an example that these inequalities cannot be improved.

38. Let x and y be two discontinuous random variables for which the possible values and probabilities are $x_1, \cdots, x_k$ and $p_1, \cdots, p_k$, $y_1, \cdots, y_l$ and $q_1, \cdots, q_l$ respectively. Let the joint probability of $x = x_i$ and $y = y_j$ be φ_{ij}, and let $k \geqq l$. Defining τ^2 by

$$\tau^2 = \frac{1}{l-1} \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(\varphi_{ij} - p_i q_j)^2}{p_i q_j},$$

show: (1) that $0 \leqq \tau^2 \leqq 1$; (2) that the necessary and sufficient condition for $\tau^2 = 0$ is that x and y are independent; (3) that the necessary and sufficient condition for $\tau^2 = 1$ is that y is completely dependent of x, i.e., that any definite value of x will with certainty be accompanied by a definite value of y.

39. Let us consider 2 players at the same bridge table, and let the random variable x denote the number of red cards of the first player, y the number of red cards of the second player. Find the correlation coefficient $\rho\{x, y\}$.

40. Let $z = (x, y)$ be an arbitrary two-dimensional random variable such that both the first- and second-order moments μ_x, μ_y, μ_{xx}, μ_{yy}, and μ_{xy} are all finite and that $\sigma_x > 0$, $\sigma_y > 0$ and $|\rho| < 1$. Show that a uniform distribution of probability mass 1 over the area enclosed by the so-called *concentration ellipse*

$$\frac{1}{1-\rho^2}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right) = 4$$

has the same first- and second-order moments as z. Next show that the two mean square regression lines are the diameters of this ellipse, conjugate to the x- and y-axis.

41. If x_1 and x_2 are two random variables both having the Cauchy distribution with the parameters μ_1, α_1 and μ_2, α_2 respectively, show that $x = x_1 + x_2$ has the Cauchy distribution with the parameters $\mu = \mu_1 + \mu_2$, $\alpha = \alpha_1 + \alpha_2$.

42. Let the probability of 1 bacteria being transformed into 2 in the time interval dt be $\lambda\, dt$. Find the probability $P_i(t)$, $i = 1, 2, 3, \cdots$, of finding i bacteria at the time t if at $t = 0$ there was 1 present. Next find $\mathfrak{M}\{x\}$ and $\sigma\{x\}$.

43. Solve problem 13 by means of generating functions.

44. An urn contains ν balls numbered 1 to ν. One ball is drawn at random, the number is observed, and the ball is put back into the urn. This experiment is repeated λ times. Find by means of generating functions the probability that the sum of the λ observed numbers has a given value.

45. Find the generating function for the binomial distribution. Apply the result to show that the sum of two binomially distributed variables is again binomially distributed in case $\theta_1 = \theta_2$.

46. Find the generating function for the Poisson distribution. Apply the result to show that the sum of two Poisson distributed variables is again Poisson distributed.

Problems to Chapters 7 through 8

47. If x is normally distributed with $\mu = 0$, find the distribution of $y = e^x$.

48. Shooting with a gun we assume the azimuth deviation, x, and the range deviation, y, to be independent and both normally distributed with the same dispersion σ. An infinitely large, horizontal target with origin in the center of impact is divided by 4 straight lines with the equations $x = 0.6745\sigma$, $x = -0.6745\sigma$, $y = 0.6745\sigma$, and $y = -0.6745\sigma$. Find the probabilities of hitting the various parts in which the target is divided.

49. N_1 beans of one sort, S_1, are mixed with N_2 beans of another sort, S_2. N_1 and N_2 are both assumed to be large numbers, and the length of both sort of beans is assumed to be normally distributed with the parameters μ_1, σ_1 and μ_2, σ_2 respectively. Find the distribution of the length of the beans in the mixture and its mean value and dispersion.

50. The height of all men at the age of conscription is assumed to be normally distributed with $\mu = 170$ cm and $\sigma = 5$ cm. At the medical examination all those are rejected whose height is less than 155 cm. Find the height distribution of the admitted men, and write down the equations for the mean value and the dispersion. Will they be smaller or larger than 170 and 5 cm respectively?

51. A random variable x is normally distributed with the parameters $\mu = 0$ and σ. We perform two independent observations of x. Let z be the largest of these two values. Show that

$$\mathfrak{M}\{z\} = \frac{\sigma}{\sqrt{\pi}}.$$

52. Let $x_1, x_2, \cdots, x_\nu$ be ν independent observations of a normally distributed random variable with $\mu = 0$ and $\sigma = 1$. The largest among these values we denote z. Find the distribution function and the probability density of z.

53. Let x, y, and z be three independent, normally distributed random variables all with the same parameters $\mu = 0$ and σ. Find the probability density for the new random variable $x^2 + y^2 + z^2$.

54. Let x and y be two independent, normally distributed random variables. Find the correlation coefficient of the two new random variables $x + y$ and $x - y$.

Problems to Chapters 10 through 11

55. Show that for the distribution

$$d\Phi = \varphi(t)\,dt = \frac{1}{2(1 - e^{-\theta})}\, e^{-|t|}\,dt, \qquad -\theta \leqq t \leqq \theta,$$

the maximum likelihood estimate of θ is the largest number among the numbers $|x_1|, \cdots, |x_n|$.

56. Let a certain event have the probability θ. If the event occurs in x out of n trials, show that x/n is the best estimate of θ.

57. In 15 measurements of the initial velocity of a projectile the following values (in meters per second) have been found:

444.9	441.0	442.1
444.2	439.5	443.8
439.8	444.1	440.2
442.6	440.9	444.9
441.8	441.3	443.7.

Compute the best estimates of the true value, of the dispersion, and of the dispersions of these estimates.

58. In 15 measurements of the latitude at Capetown the following values have been found:

−33°56′3.″48	−33°56′3.″50	−33°56′3.″50
3.″32	3.″09	2.″98
3.″07	3.″28	3.″27
3.″20	3.″30	3.″25
3.″11	3.″30	3.″27.

Compute the best estimates of the true value, of the dispersion, and of the dispersions of these estimates.

59. Shooting with a rifle the following azimuth and height deviations (in centimeters) from (0, 0) have been found:

Number	*Azimuth Deviation* x	*Height Deviation* y
1	20	−6.5
2	− 4	6.5
3	24	0
4	16	− 5
5	−10.5	1.5
6	4.5	−30
7	8	−26
8	12.5	−15
9	10	0.5
10	6.5	− 3
11	6	−13
12	1.5	− 1 .

Find the mean center of impact, the two dispersions, and the dispersions of these estimates.

60. Test whether any of the measurements of problem 59 could be suspected of being encumbered with coarse errors.

61. Test whether in problem 59 the mean center of impact deviates significantly from (0, 0).

62. Test whether in problem 59 there is any correlation between the azimuth and the height deviation.

63. In two different series of shots the following azimuth deviations (in centimeters) from (0, 0) have been found:

First Series		*Second Series*	
1	− 3.5	1	16
2	− 9	2	12
3	15.5	3	16
4	− 4	4	14.5
5	6.5	5	− 1
6	− 2.5	6	20
		7	21.5
		8	− 4
		9	9 .

Find the two estimates of the dispersion based on each separate series. Next find the best estimate as based on the total 15 measurements (cf. § 11.10).

64. Test whether in problem 63 the difference between the estimates of the dispersion from the two series is significant.

65. Test whether in problem 63 the difference between the two average values is significant.

66. The abscissae, a_1 and a_2, of two points are measured. Applying the rule of the uniform distribution of the tolerance find the allowed dispersions, σ_{a_1} and σ_{a_2}, of a_1 and a_2 when, for the distance $a = a_1 - a_2$, $\sigma_a = 0.1$ mm.

67. The density, d, of air at 0° C and 760 mm Hg is measured by finding the weight, m grams, of the air in a container having the volume V cm³ at t° C and p mm Hg:

$$d = \frac{m}{V}\frac{760}{p}\left(1 + \frac{t}{273}\right).$$

Find the value, the dispersion, and the relative dispersion of d, when

$$\begin{array}{llll} m = 2.4875\text{ g}, & V = 980.3\text{ cm}^3, & p = 741.5\text{ mm Hg}, & t = 21.4^\circ\text{ C}. \\ \sigma_m = 0.0023\text{ g}, & \sigma_V = 1.3\text{ cm}^3, & \sigma_p = 0.9\text{ mm Hg}, & \sigma_t = 0.3^\circ\text{ C}. \end{array}$$

Problems to Chapter 12

68. Show that

$$[p\varepsilon\varepsilon] = [pvv] + \underset{1n}{\mathcal{E}^*} \cdot \underset{nn}{P} \cdot \underset{nn}{H} \cdot \underset{n1}{\mathcal{E}}, \qquad \underset{nn}{H} = \underset{nn}{A} \cdot \underset{nn}{B^{-1}} \cdot \underset{nn}{A^*} \cdot \underset{nn}{P}$$

where $\underset{nn}{A}$ and $\underset{nn}{B}$ are defined in §12.3, $\underset{n1}{\mathcal{E}} = \underset{n1}{A} - \underset{n1}{L}$ is the matrix formed by the true corrections, and $\underset{nn}{H}$ is an abbreviation for the matrix mentioned (use (12.4.4)).

69. The trace (or spur), tr $\underset{nn}{A}$, of a quadratic matrix $\underset{nn}{A}$ is defined as the sum of the diagonal elements, $\text{tr } \underset{nn}{A} = \sum_{i=1}^{n} a_{ii}$. Show that tr $(\underset{nn}{A} \cdot \underset{nn}{B}) = \text{tr } (\underset{nn}{B} \cdot \underset{nn}{A})$ for two arbitrary matrices $\underset{nn}{A}$ and $\underset{nn}{B}$ for which *both* products exist.

70. Prove that

$$\sum_{i=1}^{n} p_i \sigma^2\{\bar{l}_i\} = \sigma^2 \text{ tr } (\underset{nn}{P} \cdot \underset{nn}{A} \cdot \underset{nn}{B^{-1}} \cdot \underset{nn}{A^*}),$$

where $\sigma^2\{\bar{l}_i\}$ is given in (12.8.3). Next prove (12.8.5). (Use Problem 69.)

71. Prove that

$$\mathfrak{M}\{\underset{1n}{\mathcal{E}^*} \cdot \underset{nn}{P} \cdot \underset{nn}{H} \cdot \underset{n1}{\mathcal{E}}\} = \sigma^2 \text{ tr } \underset{nn}{H}$$

(cf. problem 68). Next show that

$$\mathfrak{M}\{[pvv]\} = (n - m)\sigma^2.$$

72. To measure the distance x between the points A and B a base line AC as well as the angles $\alpha = \sphericalangle BAC$ and $\beta = \sphericalangle BCA$ have been measured. The average values of the measurements have given the numbers m, a, and b for x, α, and β respectively. Assuming $\sigma\{m\}$, $\sigma\{a\}$ and $\sigma\{b\}$ to be known, find $\sigma\{x\}$.

73. To determine the lengths of the three sides of a rectangular triangle we have measured the two sides and found the values a and b, as well as the hypotenuse and found the value c. Assuming all three measurements to be equally accurate, find the best estimates of all three sides.

74. In an equilateral triangle the three sides and the base height have been measured as

$$\{l_1, l_2, l_3, l_4\} = \{5.01 \quad 5.02 \quad 5.00 \quad 4.30\}$$

respectively. Find the best estimate of the side using adjustment by elements.

75. Solve the same problem as 74 using adjustment by correlates.

76. Test shooting a gun under increasing wind velocity perpendicular to the direction of shooting, the following values of the azimuth deviation have been found in 10 shots fired with constant time intervals:

(1) 57.2	(6) 59.8
(2) 58.0	(7) 60.4
(3) 58.1	(8) 60.0
(4) 59.1	(9) 60.0
(5) 59.3	(10) 62.2.

Assuming the azimuth deviation to be a linear function of time, $f(t) = \alpha + \beta t$, find the best estimates of the parameters α and β. (Also determine α and β graphically.)

77. For four points on a straight line, A, B, C, and D, the six distances AB, BC, CD, AC, AD, and BD have been measured:

$$\{l_1, \cdots, l_6\} = \{3.17 \quad 1.12 \quad 2.25 \quad 4.31 \quad 6.51 \quad 3.36\}.$$

Find the best estimates of the distances using adjustment by elements.

78. Solve the same problem as **77** using adjustment by correlates.

79. The number of telegrams in Germany were (in millions) in the years 1925–1934 the following:

1925	50	1930	34
1926	47	1931	27
1927	48	1932	23
1928	43	1933	22
1929	40	1934	21.

Assuming the number to be a linear function of time

$$f(t) = \alpha + \beta(t - 1925),$$

find the best estimates of the parameters α and β. (Also determine α and β graphically.)

80. The four angles of a quadrilateral have been measured as

$$\{l_1, \cdots, l_4\} = \{50°12'37'' \quad 112°17'19'' \quad 120°47'26'' \quad 76°46'18''\},$$

where l_1, l_2, l_3, l_4 are the averages of 3, 4, 2, and 2 equally accurate measurements, respectively. Find the best estimates of the angles using adjustment by elements.

81. Solve the same problem as 80 using adjustment by correlates.

82. The ordinates of three points, A, B, and C, having the abscissae 0, 1, and 2 respectively, are measured as

$$\{1.95, 2.29, 2.14\}.$$

Assuming A, B, and C to lie on one and the same circle with center on the x-axis,

$$(x - \alpha)^2 + y^2 = \rho^2,$$

find the best estimates of the parameters α and ρ.

83. The distances from a point $P = (\xi_1, \xi_2)$ to the points (0, 0), (5, 0), (0, 4), and (3, 6) have been measured as

$$\{l_1, l_2, l_3, l_4\} = \{3.60, 4.26, 2.20, 3.17\}.$$

Find the best estimates of ξ_1 and ξ_2 using adjustment by elements.

84. In a triangle ABC the sides a and b as well as the angles A and C have been measured as

$$\{a, b, A, C\} = \{35.1 \text{ cm}, 61.9 \text{ cm}, 34^\circ.1, 42^\circ.6\}.$$

Assuming that the angles have been measured with an accuracy twice as large as that of the sides find the best estimates of a, b, A, and C using adjustment by elements.

85. Solve the same problem as 84 using adjustment by correlates.

86. In measuring distances by means of a leveling instrument with horizontal wires, the distance y is a linear function of that part x of the leveling staff observed between the horizontal wires of the leveling instrument,

$$y = \alpha + \beta x.$$

Corresponding to the given values of y, 40 m, 60 m, 80 m, and 100 m, the following values of x have been measured:

$$\{0.335 \text{ m}, 0.502 \text{ m}, 0.671 \text{ m}, 0.841 \text{ m}\}.$$

Find the best estimates of the parameters α and β.

87. For 10 extra-galactic nebulae the following velocities, y (in kilometers per second), and distances, x (in millions of parsecs), have been observed:

x	y	x	y
1.20	630	9.12	4,820
1.82	890	10.97	5,230
3.31	2350	14.45	7,500
7.24	3810	22.91	11,800
6.92	4630	36.31	19,600.

Assuming y to be a linear function of x, $y = \alpha + \beta x$, find the best estimates of the parameters α and β. (Also determine α and β graphically.)

88. The pressure of a gas, p, and its volume, v, are known to be related by an equation of the form

$$pv^\gamma = \text{constant}.$$

For the data given below find the best estimate of γ by fitting a straight line to the logarithms of p and v, taking p to be the independent variable.

p (kg/cm^2)	0.5	1.0	1.5	2.0	2.5	3.0
v (liters)	1.62	1.00	0.75	0.62	0.52	0.46.

89. The average number of children per marriage, y, in Norway in 1920 for a given duration of marriage, x years, were the following

x	0–1	5–6	10–11	15–16	20–21	25–26	30–31
y	0.48	2.09	3.26	4.33	5.14	5.63	5.77.

Assuming y to be a third-order polynomial in x

$$y = \alpha_0 + \alpha_1 x^1 + \alpha_2 x^2 + \alpha_3 x^3,$$

find the best estimates of the parameters $\alpha_0, \cdots, \alpha_3$.

90. The wave lengths, λ, of the spectral lines H_α, H_β, $\cdots$ of hydrogen show a regularity found by Balmer:

$$\frac{1}{\lambda} = R\left(\frac{1}{2^2} - \frac{1}{m^2}\right), \qquad \text{i.e.,} \qquad \lambda = \frac{4}{R}\frac{m^2}{m^2 - 4},$$

in which $m = 3, 4, \cdots$ and R is called Rydberg's constant for hydrogen. From the following experimental values of λ (in Ångström units, 1 Å $= 10^{-8}$ cm) find the best estimate of R (in centimeters^{-1}):

m	λ
3	6562.79
4	4861.33
5	4340.47
6	4101.74.

REFERENCES

Chapters 1 to 9 (General theory)

E. Borel *et al.*, *Traité du calcul des probabilités et de ses applications*, Vols. 1–4, Paris, 1924–1939. (Large encyclopedia with special attention to all practical applications in economics, population statistics, biology, actuarial science, physics, ordnance, etc.)

H. Cramér, *Random Variables and Probability Distributions*, London, 1937; 120 pages. (The mathematical theory in its modern abstract form; no applications.)

E. Czuber, *Wahrscheinlichkeitsrechnung und ihre Anwendung auf Fehlerausgleichung, Statistik und Lebensversicherung*, Leipzig, 1908–1910, Vols. I–II; 410 and 470 pages.

———: "Die Entwicklung der Wahrscheinlichkeitstheorie und ihrer Anwendungen," *Jahresbericht der deutschen Mathver.*, Vol. 7, No. II, pp. 1–279, 1899. (A critical account of the history of probability and an extensive bibliography.)

M. Fréchet, "Exposé et discussion de quelques recherches récentes sur les fondements du calcul des probabilités." Théorie des probabilités, Fasc. II, p. 23, *Actualités scientifiques et industrielles*, No. 735, Paris, 1938.

———: "The Diverse Definitions of Probability," *Journal of Unified Science*, Vol. 8, p. 7, 1939.

T. C. Fry, *Probability and Its Engineering Uses*, New York, 1929; 470 pages. (Textbook especially for engineers. Gives many examples of the practical applications of probability and statistics to technology, especially telephony, and industry.)

A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin, 1933; 62 pages. (The fundamental work on the modern, abstract formulation of the theory.)

R. von Mises, *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik, Fehlertheorie und in der theoretischen Physik*, Leipzig-Wien, 1931; 574 pages. (Extensive textbook.)

———: *Probability, Statistics and Truth*, 1939; 323 pages. (A general and more philosophic discussion without mathematical details.)

Stochastic processes and limit theorems

N. Arley, *On the Theory of Stochastic Processes and Their Application to the Theory of Cosmic Radiation*, New York, 1948; 240 pages.

M. S. Bartlett, *Stochastic Processes*, Chapel Hill N.C., 1947; 95 pages. (Notes of a course given at the University of North Carolina 1946. It is understood that copies of these notes are available on request.)

S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy," *Rev. of Mod. Phys.*, Vol. 5, p. 1, 1943.

W. Feller, "Zur Theorie der Stochastischen Prozesse (Existenz und Eindeutigkeitssätze)," *Math. Ann.*, Vol. 113, p. 113, 1937.

A. Khintchine, *Asymptotische Gesetze der Wahrscheinlichkeitsrechnung*, Berlin, 1933; 77 pages.

P. Lévy, *Théorie de l'addition des variables aléatoires*, Paris, 1937; 331 pages.

O. Lundberg, *On Random Processes and Their Application to Sickness and Accident Statistics*, Uppsala, 1940; 172 pages.

J. E. Moyal, "Stochastic Processes and Statistical Physics," *J. Roy. Stat. Soc.*, Vol. B 11, 1949.

H. Wold, *A Study in the Analysis of Stationary Time Series*, Uppsala, 1938; 214 pages.

Chapter 9 (The relation of the theory of probability to experience and its practical importance)

E. Borel, "Valeur pratique et philosophie des probabilités," *Traité du calcul des probabilités*, Vol. 4, Fasc. 3, Paris, 1939; 166 pages.

C. Cranz, *Lehrbuch der Ballistik*, Bd. I, 10. Abschnitt, pp. 385–457, Berlin, 1925.

R. Fürth, *Einführung in die theoretische Physik*, Vienna, 1936; 483 pages. (Stresses the application of probability to theoretical physics.)

T. J. Hayes, *Elements of Ordnance*, New York, 1938; 715 pages.

J. Haag, "Applications du calcul des probabilités au tir," *Traité du calcul des probabilités*, Vol. 4, Fasc. 1, Paris, 1926; 182 pages. (The application of probability to ordnance.)

Chapter 10 (Application of the theory of probability to statistics)

H. Cramér, *Mathematical Methods of Statistics*, Stockholm and Princeton, 1946; 575 pages. (The most extensive and penetrating account existing. It discusses the mathematical tools for the modern, axiometic treatment of probability as well as giving an extensive account of probability itself and its applications to statistics, e.g., the theory of Fisher with rigorous mathematical conditions and proofs.)

O. L. Davies, editor, *Statistical Methods in Research and Production*, London, 1947; 292 pages. (Useful handbook without detailed proofs. Many practical examples.)

R. A. Fisher, *Statistical Methods for Research Workers*, Seventh ed., London, 1938; 356 pages and *The Design of Experiments*, Fourth ed. London 1947; 272 pages. (Much-used textbooks. Elementary and without any mathematical proofs. Give a multitude of practical examples from biology, medicine, and agricultural science.)

——— "On the Mathematical Foundations of Theoretical Statistics," *Phil. Trans. of the Roy. Soc. of London*, Vol. A 222, p. 309, 1921.

——— "Theory of Statistical Estimation," *Proc. of the Cambridge Phil. Soc.*, Vol. 22, p. 700, 1925.

——— and F. Yates, *Statistical Tables*, London, 1938; 90 pages. (Parts of the tables are also given in Fisher's book.)

A. Hald, *Statistical Theory with Engineering Applications*, New York 1952; 783 pages and tables. (Extensive textbook without advanced mathematics. Numerous numerical examples from actual practical work.)

J. C. Kapteyn and M. J. van Uven, *Skew Frequency-Curves in Biology and Statistics*, Groningen, 1916; 69 pages.

M. G. Kendall, *The Advanced Theory of Statistics*, Vols. I and II, London, 1943–1946; 457 + 521 pages. (Advanced treatment with an extensive bibliography.)

W. B. Rice, *Control Charts in Factory Management*, New York 1947; 149 pages. (Application of probability to control of quality of industrial products.)

G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Twelfth ed., London, 1940; 570 pages. (Extensive, elementary textbook.)

Chapters 11 and 12 (The application of the theory of probability to the theories of errors and of adjustment)

N. Arley, "On the Distribution of Relative Errors from a Normal Population of Errors," *Danske Vid. Selsk. Mat.-fys. Medd.*, Vol. XVIII, No. 3, 1940.

E. Czuber, *Theorie der Beobachtungsfehler*, Leipzig, 1891; 418 pages. (Critical and historical treatment.)

F. Helmert, *Die Ausgleichungsrechnung nach der Methode der kleinsten Quadrate*, Leipzig, 1907; 578 pages. (Extensive textbook with special attention to the practical applications in surveying, geodetics, physics, and the theory of measuring instruments. Many detailed numerical examples.)

C. Runge and H. König, *Vorlesungen über numerisches Rechnen*, Berlin, 1924; 372 pages.

E. T. Whittaker and G. Robinson, *The Calculus of Observations*, London, 1929; 395 pages.

§ 12.15 (Regression analysis)

M. Ezekiel, *Methods of Correlation Analysis*, Second ed., New York, 1941; 531 pages. (Comprehensive, elementary discussion of regression and correlation analysis with many practical examples.)

R. A. Fisher, *Statistical Methods for Research Workers*, Seventh ed., Chapter V, pp. 134–176, London, 1938.

A. Hald, *Statistical Theory with Engineering Applications*, Chap. 18 and 20, New York, 1952.

M. G. Kendall, *The Advanced Theory of Statistics*, Vol. II, Chapter 22, London, 1946.

G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Twelfth ed., Chapter 17, London, 1940.

Appendix 1

E. T. Whittaker and G. N. Watson, *Modern Analysis*, Fourth ed., Chapter XII, Cambridge, 1927.

Appendix 2

R. Courant and D. Hilbert, *Methoden der mathematischen Physik*, Vol. 1, Chapter 1, Berlin, 1932.

R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices*, London, 1938; 416 pages. (Elementary textbook with special attention to the practical applications and numerical calculations.)

INDEX