IX. Foundations of Probability, 375-433

DOI: 10.3792/euclid/9781429799911-9

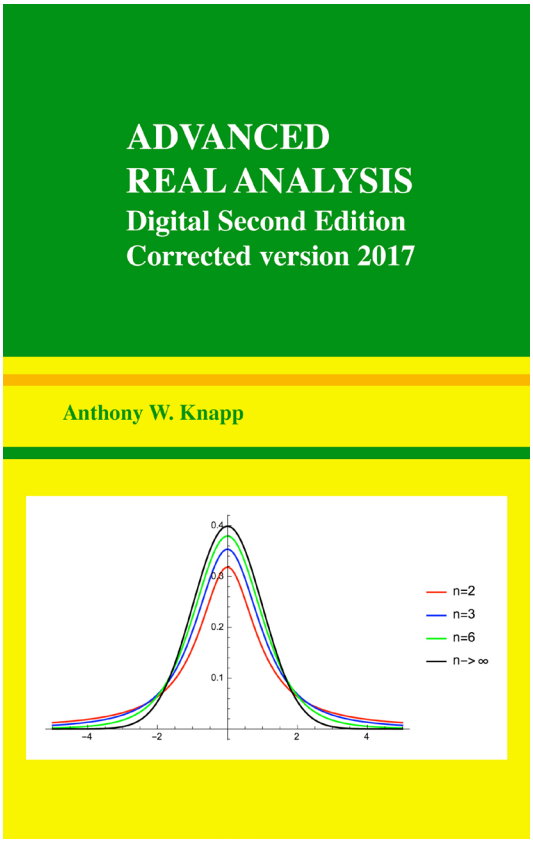
from

Advanced Real Analysis

Digital Second Edition
Corrected version 2017

Anthony W. Knapp

Full Book DOI: 10.3792/euclid/9781429799911
ISBN: 978-1-4297-9991-1



Anthony W. Knapp
81 Upper Sheep Pasture Road
East Setauket, N.Y. 11733–1729, U.S.A.
Email to: aknapp@math.stonybrook.edu
Homepage: www.math.stonybrook.edu/~aknapp

Title: Advanced Real Analysis
Cover: Normal distribution as a limit of Gosset's t distribution; see page 421.

Mathematics Subject Classification (2010): 46–01, 42–01, 43–01, 35–01, 34–01, 47–01, 58–01, 60A99, 60F05, 28C10, 42C40, 65T60.

First Edition, ISBN-13 978-0-8176-4382-9

Published by Birkhäuser Boston

Digital Second Edition, not to be sold, no ISBN

Published by the Author

CHAPTER IX

Foundations of Probability

Abstract. This chapter introduces probability theory as a system of models, based on measure theory, of some real-world phenomena. The models are measure spaces of total measure 1 and usually have certain distinguished measurable functions defined on them.

Section 1 begins by establishing the measure-theoretic framework and a short dictionary for passing back and forth between terminology in measure theory and terminology in probability theory. The latter terminology includes events, random variables, mean, probability distribution of a random variable, and joint probability distribution of several random variables. An important feature of probability is that it is possible to work with random variables without any explicit knowledge of the underlying measure space, the joint probability distributions of random variables being the objects of importance.

Section 2 introduces conditional probability and uses that to motivate the mathematical definition of independence of events. In turn, independence of events leads naturally to a definition of independent random variables. Independent random variables are of great importance in the subject and play a much larger role than their counterparts in abstract measure theory. Examples at the end of the section indicate the extent to which functions of independent random variables can remain independent. The techniques in the examples are of use in the subject of statistical inference, which is introduced in Section 10.

Section 3 states and proves the Kolmogorov Extension Theorem, a foundational result allowing one to create stochastic processes involving infinite sets of times out of data corresponding to finite subsets of those times. A special case of the theorem provides the existence of infinite sets of independent random variables with specified probability distributions.

Section 4 establishes the celebrated Strong Law of Large Numbers, which says that the Cesàro sums of a sequence of identically distributed independent random variables with finite mean converge almost everywhere to a constant random variable, the constant being the mean. This is a theorem that is vaguely known to the general public and is widely misunderstood. The proof is based on Kolmogorov's inequality.

Sections 5–8 provide background for the Central Limit Theorem, whose statement and proof are in Section 9. Section 5 discusses three successively weaker kinds of convergence for random variables—almost sure convergence, convergence in probability, and convergence in distribution. Convergence in distribution will be the appropriate kind for the Central Limit Theorem. Section 6 contains the Portmanteau Lemma, which gives some equivalent formulations of convergence in distribution, Section 7 introduces characteristic functions as Fourier transforms of probability distributions, and Section 8 proves the Lévy Continuity Theorem, which formulates convergence in distribution in terms of characteristic functions.

Section 9 contains the statement and proof of the Central Limit Theorem, followed by some simple examples. This theorem is the most celebrated result in probability theory and has many applications in mathematics and other fields.

Section 10 is a brief introduction to the subject of statistical inference, showing how the Central Limit theorem plays a role in practice through the t test of W. S. Gosset.

1. Measure-Theoretic Foundations

Although notions of probability have been around for hundreds of years, it was not until the twentieth century, with the introduction of Lebesgue integration, that the foundations of probability theory could be established in any great generality. The early work on foundations was done between 1929 and 1933 chiefly by A. N. Kolmogorov and partly by M. Fréchet.

First of all, the idea is that probability theory consists of *models* for some experiences in the real world. Second of all, these experiences are *statistical* in nature, involving repetition. Thus one attaches probability $1/2$ to the outcome of "heads" for one flip of a standard coin based on what has been observed over a period of time. One even goes so far as to attach probabilities to outcomes that one can think of repeating even if they cannot be repeated as a practical matter, such as the probability that a particular person will die from a certain kind of surgery. But one does not try to incorporate probabilities into the theory for contingencies that cannot remotely be regarded as repeatable. The philosopher R. Carnap has asked, "What is the probability that the fair coin I have just tossed has come up 'heads'?" He would insist that the answer is 0 or 1, certainly not $1/2$. Mathematical probability theory leaves his question as something for philosophers and does not address it.

The initial situation that is to be modeled is that of an experiment to be performed; the experiment may be really simple, as with a single coin toss, or it may have stages to it that may or may not be related to each other. For the moment let us suppose that the number of stages is finite; later we shall relax this condition. To fix the ideas, let us think of the outcome as a point in some Euclidean space. Forcing the outcome to be a point in a Euclidean space may not at first seem very natural for a single toss of a coin, but we can, for example, identify "heads" with 1 and "tails" with 0 in $\mathbb{R}^1$. In any case, the experiment has a certain range of conceivable outcomes, and these outcomes are to be disjoint from one another. Initially we let Ω be the set of these conceivable outcomes. If an outcome occurs when conditions belonging to a set A are satisfied, one says that the **event** A has taken place.

We imagine that probabilities have somehow been attached to the individual outcomes, and to aggregates of them, on the basis of some experimental data. Using a frequency interpretation of probability, one is led to postulate that probability in the model of this experiment is a nonnegative additive set function on some

system of subsets of Ω that assigns the value 1 to Ω itself. Without measure theory as a historical guide, one might be hard pressed to postulate complete additivity as well, but in retrospect complete additivity is not a surprising condition to impose.

At any rate, the model of the experiment within probability theory uses a measure space $(\Omega, \mathcal{A}, P)$, normally with total measure $P(\Omega)$ equal to 1, with one or more measurable functions on Ω to indicate the result of the experiment. One way of setting up $(\Omega, \mathcal{A}, P)$ is as we just did—to let Ω be the set of all possible outcomes, i.e., all possible values of the measurable functions that give the result of the experiment. Events are then simply measurable sets of outcomes, and the measure P gives the probabilities of various sets of outcomes. Yet this is not the only way, and successful work in the subject of probability theory requires a surprising indifference to the nature of the particular Ω used to model a particular experiment.

We can give a rather artificial example right now, in the context of a single toss of a standard coin, of how distinct Ω's might be used to model the same experiment, and we postpone to the last two paragraphs of this section and to the proof of Theorem 9.8 any mention of more natural situations in which one wants to allow distinct Ω's in general. The example occurs when the experiment is a single flip of a standard coin. Let us identify "heads" with the real number 1 and "tails" with the real number 0. Centuries of data and of processing the data have led to a consensus that the probabilities are to be 1/2 for each of the two possible outcomes, 1 and 0. We can model this situation by taking Ω to be the set $\{1, 0\}$ of outcomes, $\mathcal{A}$ to consist of all subsets of Ω, and P to assign weight $1/2$ to each point of Ω. The function f indicating the result of the experiment is the identity function, with $f(\omega) = 1$ if $\omega = 1$ and with $f(\omega) = 0$ if $\omega = 0$. But it would be just as good to take any other measure space $(\Omega, \mathcal{A}, P)$ with $P(\Omega) = 1$ and to suppose that there is some measurable subset A with $P(A) = 1/2$. The measurable function f modeling the experiment has $f(\omega) = 1$ if ω is in A and $f(\omega) = 0$ if not.

The problem of how to take real-world data and to extract probabilities in preparation for defining a model is outside the domain of probability theory. This involves a statistical part that obtains and processes the data, identifies levels of confidence in the accuracy of the data, and assesses the effects of errors made in obtaining the data accurately. Also it may involve making some value judgments, such as what confidence levels to treat as decisive, and such value judgments are perhaps within the domain of politicians. In addition, there is a fundamental philosophical question in whether the model, once constructed, faithfully reflects reality. This question is similar to the question of whether mathematical physics reflects the physics of the real world, but with one complication: in physics there is always the possibility that a single experimental result will disprove the model, whereas probability gives no prediction that can be disproved by a single

experimental result.

Apart from a single toss of a coin, another simple experiment whose outcome can be expressed in terms of a single real number is the selection of a "random" number from $[0, 2]$. The word "random" in this context, when not qualified in some way, insists as a matter of definition that the experiment is governed by normalized Lebesgue measure, that the probability of picking a number within a set A is the Lebesgue measure of A divided by the Lebesgue measure of $[0, 2]$. If we take Ω to be $[0, 2]$, $\mathcal{A}$ to be the Borel sets, and P to be $\frac{1}{2}\,dx$ and if we use the identity function as the measurable function telling the outcome, then we have completely established a model.

The theory needed for setting up a model that incorporates given probabilities is normally not so readily at hand, since one is quite often interested potentially in infinitely many stages to an experiment and the given data concern only finitely many stages at a time. In many cases of this kind, one invokes a fundamental theorem of Kolmogorov to set up a measure space that can allow the set of distinguished measurable functions to be infinite in number. We shall state and prove this theorem in Section 3.

In the meantime let us take the measure space $(\Omega, \mathcal{A}, P)$ with $P(\Omega) = 1$ as given to us. We refer to $(\Omega, \mathcal{A}, P)$ or simply (Ω, P) as a **probability space**. Probability theory has its own terminology. An **event** is a measurable set, thus a set in the σ-algebra $\mathcal{A}$. One speaks of the "probability of an event," which means the P measure of the set. The language used for an event is often slightly different from the ordinary way of defining a set. With the random-number example above, one might well speak of the probability of the "event that the random number lies in $[1/2, 1]$" when a more literal description is that the event *is* $[1/2, 1]$. It is not a large point. The probability in either case, of course, is $1/4$.

Let A and B be events. The event $A \cap B$ is the simultaneous occurrence of A and B. The event $A \cup B$ is the event that at least one of A and B occurs. The event A^c is the nonoccurrence of the event A. If $A = \varnothing$, event A is impossible; if $A = \Omega$, event A must occur. Containment $B \subseteq A$ means that from the occurrence of event B logically follows the occurrence of event A. Two events A and B are incompatible if $A \cap B = \varnothing$. A set-theoretic partitioning C of Ω as a disjoint union $\Omega = \bigcup_{k=1}^{n} A_k$ corresponds to an experiment C consisting of determining which of the events $A_1, \ldots, A_n$ occurs. And so on.

A **random variable** is a real-valued measurable function on Ω. With the random-number example, a particular random variable is the number selected. This is the function f that associates the real number ω to the member ω of the space Ω. The word "random" in the name "random variable" refers to the fact that its value depends on which possibility in Ω is under consideration. Some latitude needs to be made in the definition of measurable function to allow a function taking on values "heads" and "tails" to be a random variable, but this point will

not be important for our purposes.[1] As we shall see, the random variables that yield the result of the defining experiment of a probability model are, in a number of important cases, coordinate functions on a set Ω given as a product, and random variables are often indicated by letters like x suitable for coordinates.[2]

The **mean** or **expectation** or **expected value** $E(x)$ of the random variable x is motivated by a computation in the especially simple case that Ω contains finitely many outcomes/points and $P(A)$ is computed for an event by adding the weights attached to the outcomes ω of A. If ω is an outcome, the value of x at ω is $x(\omega)$, and this outcome occurs with probability $P(\{\omega\})$. Summing over all outcomes, we obtain $\sum_{\omega\in\Omega} x(\omega)P(\{\omega\})$ as a reasonable notion of the expected value. This sum suggests a Lebesgue integral, and accordingly the definition in the general case is that $E(x) = \int_\Omega x(\omega)\,dP(\omega)$. Probabilists say that $E(x)$ **exists** if x is *integrable*; cases in which the Lebesgue integral exists and is infinite are excluded.

There is a second way of computing the mean. When Ω is a finite set as above, we can group all the terms in $\sum_{\omega\in\Omega} x(\omega)P(\{\omega\})$ for which $x(\omega)$ takes a particular value c and then sum on c. The regrouped value of the sum is $\sum_c cP(\{\omega \mid x(\omega) = c\})$. The corresponding formula in the general case involves the **probability distribution** of x, the Stieltjes measure μ_x on the Borel sets of the line $\mathbb{R}$ defined by

$$\mu_x(A) = P(\{\omega \in \Omega \mid x(\omega) \in A\}).$$

The name "distribution" is traditional in probability theory to emphasize the way in which mass has been spread in some fashion, and the adjective "probability" refers to the fact that this measure has total mass $\mu_x(\mathbb{R}) = P(\Omega) = 1$. Although Stieltjes measures are indeed distributions in the sense of Chapter V, it is not at all helpful to think of them in this way in probability theory. Thus we shall usually retain the adjective "probability" to head off any confusion.

The notion of μ_x, but not the name, was introduced in Section VI.10 of *Basic*. The formula for the mean in terms of the probability distribution of x is $E(x) = \int_{\mathbb{R}} x\,d\mu_x$; the justification for this formula lies in the following proposition, which was proved in *Basic* as Proposition 6.56a and which we re-prove here.

Proposition 9.1. If $x : \Omega \to \mathbb{R}$ is a random variable on a probability space

[1] We return to this point in Section 3, where it will influence the hypotheses of the fundamental theorem of Kolmogorov.

[2] In his book *Measure Theory* Doob writes on p. 179, "An attentive reader will observe ... that in other chapters a function is f or g, and so on, whereas in this chapter [on probability] a function is more likely to be x or y, and so on, at the other end of the alphabet. This difference is traditional, and is one of the principal features that distinguishes probability from the rest of measure theory."

(Ω, P) and if μ_x is the probability distribution of x, then

$$\int_\Omega \Phi(x(\omega))\, dP(\omega) = \int_{\mathbb{R}} \Phi(t)\, d\mu_x(t)$$

for every nonnegative Borel measurable function $\Phi : \mathbb{R} \to \mathbb{R}$. The formula extends linearly to the case in which the condition "nonnegative" on Φ is dropped if the integrals for $\Phi^+ = \max(\Phi, 0)$ and $\Phi^- = -\min(\Phi, 0)$ are both finite. It extends to complex-valued Φ if the integral for $|\Phi|$ is finite.

PROOF. When Φ is the indicator function I_A of a Borel set A of $\mathbb{R}$, the two sides of the identity are $P(x^{-1}(A))$ and $\mu_x(A)$, and these are equal by definition of μ_x. We can pass to nonnegative simple functions by linearity and then to general nonnegative Borel measurable functions Φ by monotone convergence. □

The qualitative conclusion of Proposition 9.1 is by itself important: the mean of any function of a random variable can be computed in terms of the probability distribution of the random variable—without reference to the underlying measure space Ω.

The expression for $E(x)$ arising from Proposition 9.1 can often be written as a "Stieltjes integral," which is a simple generalization of the Riemann integral,[3] and thus the proposition in principle gives a way of computing means without Lebesgue integration.[4]

Instead of working with the Stieltjes measure μ_x, one can work with an associated monotone function on $\mathbb{R}$. The particular monotone function used by probabilists is the **cumulative distribution function** of x, defined by

$$F_x(t) = \mu_x((-\infty, t]).$$

The cumulative distribution function of x differs only by the additive constant $\mu_x((-\infty, 0])$ from the distribution function introduced in Section VI.8 of *Basic*; the value of the latter monotone function at t was

$$\begin{cases} -\mu((x, 0]) & \text{if } x \le 0 \\ \mu((0, x]) & \text{if } x \ge 0. \end{cases}$$

When the probability measure μ_x is absolutely continuous with respect to Lebesgue measure, we can write $\mu_x = f_x(t)\, dt$ for a function f_x by the Radon–Nikodym Theorem.[5] Such a function f_x, which is determined up to sets of

[3] Stieltjes integration is developed briefly in the problems at the end of Chapter III of *Basic*.

[4] Consequently the resulting formula for means is handy pedagogically and is often exploited in elementary probability books.

[5] Corollary 7.10 or Theorem 9.16 of *Basic*.

measure 0, is called the **density** of the random variable x. In terms of monotone functions, a density exists if and only if the cumulative distribution function is absolutely continuous (for example, when it has a continuous derivative), and in this case the density is the pointwise derivative a.e. of the cumulative distribution function. If x has a density f_x, the formula for the mean becomes $E(x) = \int_{\mathbb{R}} t f_x(t)\, dt$; this conclusion is just Proposition 9.1 for the Borel function $\Phi(t) = t$. More generally, $E(\Phi(x)) = \int_{\mathbb{R}} \Phi(t) f_x(t)\, dt$ for any Φ as in Proposition 9.1.

A set of random variables is said to be **identically distributed** if all of them have the same Stieltjes measure as probability distribution. As a consequence of Proposition 9.1, identically distributed random variables have the same mean. We shall make serious use of identically distributed random variables starting in Section 4.

Although Proposition 9.1 allows us to compute the mean of any Borel function of a random variable in terms of the probability distribution of the random variable, it does not help us when we have to deal with more than one random variable. The appropriate device for more than one random variable is a "joint probability distribution." If $x_1, \ldots, x_N$ are random variables, define, for each Borel set A in $\mathbb{R}^N$,

$$\mu_{x_1,\ldots,x_N}(A) = P\big(\{\omega \in \Omega \mid \big(x_1(\omega), \ldots, x_N(\omega)\big) \in A\}\big).$$

Then $\mu_{x_1,\ldots,x_N}$ is a Borel measure on $\mathbb{R}^N$ with $\mu_{x_1,\ldots,x_N}(\mathbb{R}^N) = 1$. It is called the **joint probability distribution** of $x_1, \ldots, x_N$. Referring to the definition, we see that we can obtain the joint probability distribution of a subset of $x_1, \ldots, x_N$ by dropping the relevant variables: for example, dropping x_N enables us to pass from the joint probability distribution of $x_1, \ldots, x_N$ to the joint probability distribution of $x_1, \ldots, x_{N-1}$, the formula being

$$\mu_{x_1,\ldots,x_{N-1}}(B) = \mu_{x_1,\ldots,x_N}(B \times \mathbb{R}).$$

Proposition 9.2. If $x_1, \ldots, x_N$ are random variables on a probability space (Ω, P) and if $\mu_{x_1,\ldots,x_N}$ is their joint probability distribution, then

$$\int_\Omega \Phi(x_1(\omega), \ldots, x_N(\omega))\, dP(\omega) = \int_{\mathbb{R}^N} \Phi(t_1, \ldots, t_N)\, d\mu_{x_1,\ldots,x_N}(t_1, \ldots, t_N)$$

for every nonnegative Borel measurable function $\Phi : \mathbb{R}^N \to \mathbb{R}$. The formula extends linearly to the case in which the condition "nonnegative" on Φ is dropped if the integrals for $\Phi^+ = \max(\Phi, 0)$ and $\Phi^- = -\min(\Phi, 0)$ are both finite. It extends to complex-valued Φ if the integral for $|\Phi|$ is finite.

PROOF. In (a), when Φ is the indicator function I_A of a Borel set A of $\mathbb{R}^N$, the two sides of the identity are $P((x_1, \ldots, x_N)^{-1}(A))$ and $\mu_{x_1,\ldots,x_N}(A)$, and these

are equal by definition of $\mu_{x_1,\dots,x_N}$. We can pass to nonnegative simple functions by linearity and then to general nonnegative Borel measurable functions Φ by monotone convergence. □

As with Proposition 9.1, the qualitative conclusion of Proposition 9.2 is by itself important: the mean of any function of N random variables can be computed in terms of their joint probability distribution—without reference to the underlying measure space Ω. For example the product of the N random variables is a function of them, and therefore

$$E(x_1 \cdots x_N) = \int_{\mathbb{R}^N} t_1 \cdots t_N \, d\mu_{x_1,\dots,x_N}(t_1, \dots, t_N).$$

When the probability measure $\mu_{x_1,\dots,x_N}$ is absolutely continuous with respect to Lebesgue measure, we can write $\mu_{x_1,\dots,x_N} = f_{x_1,\dots,x_N}(t)\,dt$ for a function $f_{x_1,\dots,x_N}$ by the Radon–Nikodym Theorem.[6] Such a function $f_{x_1,\dots,x_n}$, which is determined up to sets of measure 0, is called the **joint probability density** of the random variables $x_1, \dots, x_N$.

The possibility of making such computations without explicitly using Ω has the effect of changing the emphasis in the subject. Often it is not that one is given such-and-such probability space and such-and-such random variables on it. Instead, one is given some random variables and, if not their precise joint probability distribution, at least some properties of it. Accordingly, we can ask, What Borel measures μ on $\mathbb{R}^N$ with $\mu(\mathbb{R}^N) = 1$ are joint probability distributions of some family $x_1, \dots, x_N$ of N random variables on some probability space (Ω, P)?

The answer is, *all* Borel measures μ with $\mu(\mathbb{R}^N) = 1$. In fact, we have only to take $(\Omega, P) = (\mathbb{R}^N, \mu)$ and let x_j be the j^{th} coordinate function $x_j(\omega_1, \dots, \omega_N) = \omega_j$ on $\mathbb{R}^N$. Substituting into the definition of joint probability distribution, we see that the value of the joint probability distribution $\mu_{x_1,\dots,x_N}$ on a Borel set A in $\mathbb{R}^N$ is

$$\begin{aligned}\mu_{x_1,\dots,x_N}(A) &= \mu(\{\omega \in \mathbb{R}^N \mid (x_1(\omega), \dots, x_N(\omega)) \in A\}) \\ &= \mu(\{\omega \in \mathbb{R}^N \mid (\omega_1, \dots, \omega_N) \in A\}) = \mu(A).\end{aligned}$$

Thus $\mu_{x_1,\dots,x_N}$ equals the given measure μ.

Even for $N = 1$, this conclusion is useful. In the proof of the Central Limit Theorem later in this chapter, we shall encounter the "normal distribution on $\mathbb{R}^1$ with mean 0 and variance σ^2." This is the Stieltjes measure μ on $\mathbb{R}$ defined by

$$\mu(A) = \frac{1}{\sigma\sqrt{2\pi}} \int_A e^{-u^2/(2\sigma^2)}\,du$$

[6] Theorem 9.16 of *Basic*.

for every Borel set A, i.e., the absolutely continuous Stieltjes measure with density $(\sigma\sqrt{2\pi})^{-1}\, e^{-u^2/(2\sigma^2)}$; it has $\mu(\mathbb{R}) = 1$. The above remarks show how to define a random variable with this particular probability distribution: the underlying space is $\Omega = \mathbb{R}$, the underlying probability measure is this μ, and the random variable is the coordinate function x on $\mathbb{R}$.

2. Independent Random Variables

The notion of independence of events in probability theory is a matter of definition, but the definition tries to capture the intuition that one might attach to the term. Thus one seeks a mathematical condition saying that a set of attributes determining a first event has no influence on a second event and vice versa. Kolmogorov writes,[7]

> Historically, the independence of experiments and random variables represents the very mathematical concept that has given the theory of probability its peculiar stamp. The classical work of LaPlace, Poisson, Tchebychev, Liapounov, Mises, and Bernstein is actually dedicated to the fundamental investigation of series of independent random variables. . . . We thus see, in the concept of independence, at least the germ of the peculiar type of problem in probability theory. . . . In consequence, one of the most important problems in the philosophy of the natural sciences is—in addition to the well-known one regarding the essence of the concept of probability itself—to make precise the premises which would make it possible to regard any given real events as independent.

The path to discovering the mathematical condition that captures independence of events begins with "conditional probability." Let A and B be two events, and assume that $P(B) > 0$. Think of A as a variable. The **conditional probability** of A given B, written $P(A \mid B)$, is to be a new probability measure, as A varies, and is to be a version of P adjusted to take into account that B happens. These words are interpreted to mean that a normalization is called for, and the corresponding definition is therefore

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

In measure-theoretic terms, we pass from the measure space $(\Omega, \mathcal{A}, P)$ to the measure space $\big(B,\ \mathcal{A} \cap B,\ P((\,\cdot\,) \cap B)\big/P(B)\big)$. Conditional probabilities $P(A \mid B)$ are left undefined when $P(B) = 0$.

[7] In his *Foundations of the Theory of Probability*, second English edition, pp. 8–9.

The intuition concerning independence of A and B is that the occurrence of B is not to influence the probability of A. Thus two events A and B are to be independent, at least when $P(B) > 0$, if $P(A) = P(A \mid B)$. This condition initially looks asymmetric, but if we substitute the definition of conditional probability, we find that the condition is $P(A) = \frac{P(A \cap B)}{P(B)}$, hence that

$$P(A \cap B) = P(A)P(B).$$

This condition is symmetric, and it allows us to drop the assumption that $P(B) > 0$. We therefore define the events A and B to be **independent** if $P(A \cap B) = P(A)P(B)$.

As the quotation above from Kolmogorov indicates, the question of the extent to which this definition of independence captures from nature our intuition for what the term should mean is a deep fundamental problem in the philosophy of science. We shall not address it further.

But a word of caution is appropriate. The assumption of mathematical independence carries with it far-reaching consequences, and it is not to be treated lightly. Members of the public all too frequently assume independence without sufficient evidence for it. Here are two examples that made national news in the first decade of the twenty-first century.

EXAMPLES.

(1) In the murder trial of a certain sports celebrity, a criminalist presented evidence that three characteristics of some of the blood at the scene matched the defendant's blood, and the question was to quantify the likelihood of this match if the defendant was not the murderer. Two of the three characteristics amounted to the usual blood type and Rh factor, and the criminalist said that half the people in the population had blood with these characteristics. The third characteristic was something more unusual, and he asserted that only 4% of the population had blood with this characteristic. He concluded that only 2% of the population had blood for which these three characteristics matched those in the defendant's blood and the blood at the scene. The defense attorney jumped on the criminalist, asking how he arrived at the 2% figure, and received a confirmation that the criminalist had simply multiplied the probability .5 for the blood type and Rh factor by the .04 for the third characteristic. Upon being questioned further, the criminalist acknowledged that he had multiplied the probabilities because he could not see that these characteristics had anything to do with each other. The defense attorney elicited a further acknowledgment that the criminalist was aware of no studies of the joint probability distribution. The criminalist's testimony was thus discredited, and the jurors could ignore it. What the criminalist could have said, but did not, was that anyway at most 4% of the population had blood with

those three characteristics because of that third characteristic alone; that assertion would not have required any independence.

(2) In the 2004 presidential election, some malfunctions involving electronic voting machines occurred in three states in a particular way that seemed to favor one of the two main candidates. One national commentator who pursued this story rounded up an expert who examined closely what happened in one of the states and came up with a rather small probability of about .1 for the malfunction to have been a matter of pure chance. Seeing that the three states were widely separated geographically and that communication between officials of the different states on Election Day was unlikely, the commentator apparently concluded in his mind that the three events were independent. So he multiplied the probabilities and announced to the public that the probability of this malfunction in all three states on the basis of pure chance was a decisively small .001. What he ignored was that the machines in the three states were all made by the same company; so the assumption of independence was doubtful.

Of more importance for our purposes than independence of events is the notion of independence of random variables. Tentatively let us say that two random variables x and y on a probability space (Ω, P) are defined to be independent if $\{x(\omega) \in A\}$ and $\{y(\omega) \in B\}$ are independent events for every pair of Borel subsets A and B of $\mathbb{R}$. Substituting the definition of independent events, we see that the condition is that

$$P(\{\omega \mid (x(\omega), y(\omega)) \in A \times B\}) = P(\{\omega \mid x(\omega) \in A\})P(\{\omega \mid y(\omega) \in B\})$$

for every pair of Borel subsets of $\mathbb{R}$. We can rewrite this condition in terms of their probability distributions as

$$\mu_{x,y}(A \times B) = \mu_x(A)\mu_y(B).$$

In other words, the measure $\mu_{x,y}$ on $\mathbb{R}^2$ agrees with the product measure $\mu_x \times \mu_y$ on measurable rectangles. By Proposition 5.45 of *Basic*, the two measures must then agree on all Borel sets of $\mathbb{R}^2$. Conversely if the two measures agree on all Borel sets of $\mathbb{R}^2$, then they agree on all measurable rectangles. We therefore adopt the following definition: two random variables x and y on a probability space (Ω, P) are **independent** if their joint probability distribution is the product of their individual probability distributions, i.e., if $\mu_{x,y} = \mu_x \times \mu_y$.

One can go through a similar analysis, starting from conditional probability involving N events, and be led to a similar result for N random variables. The upshot is that N random variables $x_1, \ldots, x_N$ on a probability space (Ω, P) are defined to be **independent** if their joint probability distribution $\mu_{x_1,\ldots,x_N}$ is the N-fold product of the individual probability distributions $\mu_{x_1}, \ldots, \mu_{x_N}$. An

infinite collection of random variables is said to be **independent** if every finite subcollection of them is independent.

We can ask whether arbitrarily large finite numbers of independent random variables exist on some probability space with specified probability distributions, and the answer is "yes." This question is a special case of the one near the end of Section 1. If we are given N Borel measures $\mu_1, \dots, \mu_N$ on $\mathbb{R}$ and we seek independent random variables with these measures as their respective individual probability distributions, we form the product measure $\mu = \mu_1 \times \cdots \times \mu_N$. Then the observation at the end of Section 1 shows us that if we take $(\mathbb{R}^N, \mu)$ as a probability space and if we define N random variables on $\mathbb{R}^N$ to be the N coordinate functions, then the N random variables have μ as joint probability distribution. Since μ is a product, the random variables are independent.

The question is more subtle if asked about infinitely many independent random variables. If, for example, we are given an infinite sequence of Borel measures on $\mathbb{R}$, we do not yet have tools for obtaining a probability space with a sequence of independent random variables having those individual probability distributions.[8] We can handle an arbitrarily large finite number, and we need a way to pass to the limit. The passage to the limit for this situation is the simplest nontrivial application of the fundamental theorem of Kolmogorov that was mentioned in Section 1. The theorem will be stated and proved in Section 3.

We conclude this section with two propositions and some examples concerning independence.

Proposition 9.3. If $x_1, \dots, x_N$ are independent random variables on a probability space, then $E(x_1 \cdots x_N) = E(x_1) \cdots E(x_N)$.

PROOF. If $\mu_{x_1,\dots,x_N}$ is the joint probability distribution of $x_1, \dots, x_N$, then it was observed after Proposition 9.2 that

$$E(x_1 \cdots x_N) = \int_{\mathbb{R}^N} t_1 \cdots t_N \, d\mu_{x_1,\dots,x_N}(t_1, \dots, t_n). \tag{$*$}$$

The independence means that $d\mu_{x_1,\dots,x_N}(t_1, \dots, t_n) = d\mu_{x_1}(t_1) \cdots d\mu_{x_N}(t_N)$. Then the integral on the right side of $(*)$ splits as the product of N integrals, the j^{th} factor being $\int_{\mathbb{R}} t_j \, d\mu_{x_j}(t_j)$. This j^{th} factor equals $E(x_j)$, and the proposition follows. □

Proposition 9.4. Let

$$x_1, \dots, x_{k_1}, x_{k_1+1}, \dots, x_{k_2}, x_{k_2+1}, \dots, x_{k_3}, \quad \dots, \quad x_{k_{m-1}+1}, \dots, x_{k_m}$$

[8] There is one trivial case that we can already handle. An arbitrary set of constant random variables can always be adjoined to an independent set, and the independence will persist for the enlarged set.

be k_m independent random variables on a probability space, define $k_0 = 0$, and suppose that $F_j : \mathbb{R}^{k_j - k_{j-1}} \to \mathbb{R}$ is a Borel function for each j with $1 \le j \le m$. Then the m random variables $F_j(x_{k_{j-1}+1}, \dots, x_{k_j})$ are independent.

REMARKS. That is, functions of disjoint subsets of a set of independent random variables are independent.

PROOF. Put $y_j = (x_{k_{j-1}+1}, \dots, x_{k_j})$, and define $y = (y_1, \dots, y_m)$ and $F = (F_1, \dots, F_m)$. Let $\mathbb{R}_j$ be the copy of $\mathbb{R}^{k_j - k_{j-1}}$ corresponding to variables numbered $k_{j-1} + 1$ through k_j, and regard the probability distribution $\mu_{F_j(y_j)}$ of F_j as a measure on $\mathbb{R}_j$. What needs proof is that

$$\mu_{F(y)} = \mu_{F_1(y_1)} \times \cdots \times \mu_{F_m(y_m)}. \qquad (*)$$

Both sides of this expression are Borel measures on $\mathbb{R}^{k_m}$. On any product set $A = A_1 \times \cdots \times A_m$, where A_j is a Borel subset of $\mathbb{R}_j$, we have

$$\begin{aligned}
\mu_{F(y)}(A) &= P(\{\omega \mid F(y(\omega)) \in A\}) \\
&= P(\{\omega \mid F_j(y_j(\omega)) \in A_j \text{ for all } j\}) \\
&= P(\{\omega \mid y_j(\omega) \in F_j^{-1}(A_j) \text{ for all } j\}) \\
&= \textstyle\prod_{j=1}^m P(\{\omega \mid y_j(\omega) \in F_j^{-1}(A_j)\}) \quad \text{by the assumed independence} \\
&= \textstyle\prod_{j=1}^m P(\{\omega \mid F_j(y_j)(\omega) \in A_j\}) \\
&= \textstyle\prod_{j=1}^m \mu_{F_j(y_j)}(A_j).
\end{aligned}$$

Consequently the two sides of $(*)$ are equal on all Borel sets. □

Now let us come to some examples. Proposition 9.4 is a useful tool for generating independent random variables, as Examples 1 and 2 will show. On the other hand, independence of random variables is not as robust a notion as one might hope, according to Example 3a. Examples 3b and 3c are motivated by Example 3a and develop a change-of-variables technique that is useful in Section 10. Example 4 is a complement to Example 3a, showing that sometimes independence occurs in new random variables defined in terms of other random variables even when Examples 1 and 2 do not apply; this situation will be of critical importance in Section 10.

EXAMPLES.

(1) If $x_1, x_2, \dots, x_N$ are independent random variables and $F_1, F_2, \dots, F_N$ are Borel functions on $\mathbb{R}^1$, then $F_1(x_1), F_2(x_2), \dots, F_N(x_N)$ are independent random variables.

(2) If $x_1, \dots, x_N$ are independent random variables and if $s_j = x_1 + \cdots + x_j$, then the two random variables s_j and $s_N - s_j$ are independent because s_j depends only on $x_1, \dots, x_j$ and $s_N - s_j$ depends only on $x_{j+1}, \dots, x_N$.

(3) Suppose that two independent random variables x_1 and x_2 are given, and suppose that we form two new random variables $y_1 = f_1(x_1, x_2)$ and $y_2 = f_2(x_1, x_2)$. Let us focus on what happens under the change of variables. For simplicity suppose that the vector-valued function $f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$ is smooth and is invertible with smooth inverse given by $g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$. Suppose also that x_1 and x_2 both have densities: $\mu_{x_1} = h_1(t_1)\,dt_1$ and $\mu_{x_2} = h_2(t_2)\,dt_2$. The joint probability distribution of x_1 and x_2 is $\mu_{x_1,x_2} = h_1(t_1)h_2(t_2)\,dt_1\,dt_2$ because of the assumed independence, and thus x_1 and x_2 have $h_1(t1)h_2(t_2)$ as joint probability density. Proposition 9.2 shows that

$$\int_\Omega \Phi(x_1(\omega), x_2(\omega))\,dP(\omega) = \int_{\mathbb{R}^2} \Phi(t_1, t_2)h_1(t_1)h_2(t_2)\,dt_1\,dt_2. \tag{$*$}$$

for every nonnegative Borel function Φ. We shall apply this formula in three situations.

(3a) The first question is whether y_1 and y_2 are independent. For testing independence of y_1 and y_2, let Φ be the composition $\Phi = I_{A\times B} \circ f$, where $I_{A\times B}$ is the indicator function of the product set. The left side of $(*)$ simplifies to $\mu_{y_1,y_2}(A \times B)$, and we evaluate the right side by making the change of variables $\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = g\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$; the tool is Theorem 6.32 of *Basic*. The right side equals

$$\int_{A\times B} h_1(g_1(u_1, u_2))h_2(g_2(u_1, u_2))\left|\det\left[\tfrac{\partial t_i}{\partial u_j}\right]\right| du_1\,du_2, \tag{$**$}$$

and the question is whether this expression is a product $\nu_1(A)\nu_2(B)$ for Stieltjes measures ν_1 and ν_2. Essentially this is the question whether the integrand is the product of a function of u_1 and a function of u_2. A fairly simple case is that f is a specific linear function from $\mathbb{R}^2$ to $\mathbb{R}^2$, say $f(x_1, x_2) = (x_1 + x_2, x_1 - x_2)$ with inverse $g(y_1, y_2) = (\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2))$. Then $\det\left[\frac{\partial t_i}{\partial u_j}\right]$ is the constant function $-\frac{1}{2}$, and we are to consider an integrand of the form

$$\tfrac{1}{2}h_1(\tfrac{1}{2}(u_1 + u_2))h_2(\tfrac{1}{2}(u_1 - u_2)).$$

Without some special assumption on h_1 and h_2, this integrand has little chance of being the product of a function of u_1 by a function of u_2. Thus y_1 and y_2 will fail to be independent without special additional assumptions.

(3b) The second question is how to deduce from $(*)$ information about the probability distribution of a real-valued function of two variables. Let us take the

function $(x_1, x_2) \mapsto x_1 + x_2$ as an example. The question is to find the probability distribution of $x_1 + x_2$ when x_1 and x_2 are known to be independent. The device is to view $(x_1, x_2) \mapsto x_1 + x_2$ as one coordinate of a change of variables. We can take the other coordinate to be $(x_1, x_2) \mapsto x_2$, so that we are considering the change of variables $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} t_1+t_2 \\ t_2 \end{pmatrix}$ with inverse $\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \begin{pmatrix} u_1-u_2 \\ u_2 \end{pmatrix}$. Let Φ be the composition of this function followed by $I_{A\times\mathbb{R}}$. Formulas $(*)$ and $(**)$ give

$$\begin{aligned}\int_\Omega I_{A\times\mathbb{R}}\begin{pmatrix} x_1(\omega)+x_2(\omega) \\ x_2(\omega)\end{pmatrix} dP(\omega) &= \int_{\mathbb{R}^2} I_{a\times\mathbb{R}}\begin{pmatrix} t_1+t_2 \\ t_2\end{pmatrix} h_1(t_1)h_2(t_2)\,dt_1\,dt_2 \\ &= \int_{\mathbb{R}^2} I_{A\times\mathbb{R}}(u_1,u_2)h_1(u_1-u_2)h_2(u_2)\,du_1\,du_2 \\ &= \int_{\mathbb{R}} I_A(u_1)\Big(\int_{\mathbb{R}} h_1(u_1-u_2)h_2(u_2)\,du_2\Big)\,du_1.\end{aligned}$$

The left side is just $\int_A (x_1(\omega) + x_2(\omega))\,dP(\omega)$, and thus this equality says that the density of the sum of two random variables x_1 and x_2 is the convolution of the separate densities of x_1 and x_2.

(3c) The third question is what happens when $(x_1, x_2) \mapsto x_1 + x_2$ in (3b) is replaced by some more general scalar-valued function $(x_1, x_2) \mapsto \varphi(x_1, x_2)$ with φ smooth. Going over what happened in (3b), we see that we can certainly embed this in a smooth change of variables if the partial derivative of φ in the first variable is everywhere positive. The change of variables is then $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \varphi(t_1,t_2) \\ t_2 \end{pmatrix}$, and the Jacobian determinant is $\frac{\partial\varphi}{\partial t_1} \neq 0$. We can invert by means of either direct computation or the Inverse Function Theorem[9] and integrate out one variable just as in (3b). We shall use this technique in Section 10.

(4) We now exhibit an assumption that succeeds in yielding independence in Example 3. Suppose that n independent random variables $x_1, \ldots, x_n$ are given, and suppose that new random variables $y_1, \ldots, y_n$ are formed by a linear function, specifically that

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

for an invertible square matrix A. Suppose further that $x_1, \ldots, x_n$ have densities given by a quadratic exponential independent of j: $\mu_{x_j} = ce^{-ax_j^2}\,dx_j$ for all j, where a is positive and c is chosen to make μ_{x_j} have total mass 1. Using the technique of Example 3, let us ask whether $y_1, \ldots, y_n$ are independent. In Example 3 we have $h_j(t_j) = ce^{-at_j^2}$ and

$$\mu_{x_1,\ldots,x_n} = c^n \textstyle\prod_j e^{-at_j^2}\,dt_j = c^n e^{-at\cdot t}\,dt_1\cdots dt_n, \quad \text{where } t = (t_1, \ldots, t_n).$$

[9] Theorem 3.17 of *Basic*.

We write $u = (u_1, \dots, u_n) = A^{-1}t$ and substitute into $(**)$ of Example 3. The factor $\left|\det\left[\frac{\partial t_i}{\partial u_j}\right]\right|$ is a positive constant p, and thus

$$\mu_{y_1,\dots,y_n} = c^n p e^{-a(A^{-1}u)\cdot(A^{-1}u)}\, du_1 \cdots du_n = c^n p e^{-a(A^{-1})^{\mathrm{tr}}(A^{-1}u)\cdot u}\, du_1 \cdots du_n.$$

Because of the transformation property of the exponential function, the coefficient function on the right side is a product of functions of each variable if $(A^{-1})^{\mathrm{tr}}(A^{-1})$ is a diagonal matrix. As a result, the transformed random variables are independent if $(A^{-1})^{\mathrm{tr}}(A^{-1})$ is a diagonal matrix. An example of a nonsquare matrix with this property is $A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$. We shall make use of this idea in Section 10.

3. Kolmogorov Extension Theorem

The problem addressed by the Kolmogorov theorem is the setting up of a "stochastic process," a notion that will be defined presently. Many stochastic processes have a time variable in them, which can be discrete or continuous. The process has a set S of "states," which can be a finite set, a countably infinite set, or a suitably nice uncountable set. It will be sufficient generality for our purposes that the set of states be realizable as a subset of a Euclidean space, the measurable subsets of states being the intersection of S with the Borel sets of the Euclidean space. The defining measurable functions tell the state at each instant of time. Accordingly, one might want to enlarge the definition of random variable to allow the range to contain S. But we shall not do so, instead referring to "measurable functions" in the appropriate places rather than random variables.

Let us give one example of a stochastic process with discrete time and another with continuous time, with particular attention to the passage to the limit that is needed in order to have a probability model realizing the stochastic process.

In the example with discrete time, we shall assume also that the state space S is countable. The probabilistic interpretation of the situation visualizes the process as moving from state to state as time advances through the positive integers, with probabilities depending on the complete past history but not the future; but this interpretation will not be important for us. Let us consider the analysis. In the n^{th} finite approximation $(\Omega_n, \mathcal{A}_n, P_n)$ for $n \geq 1$, the set Ω_n is countable and consists of all ordered n-tuples of members of S, while $\mathcal{A}_n$ is the set of all subsets of Ω_n. The measure P_n is determined by assigning a nonnegative weight to each member of Ω_n, the sum of all the weights being 1. As n varies, a consistency condition is to be satisfied: the sum over S of all the weights in Ω_{n+1} of the $(n+1)$-tuples that start with a particular n-tuple is the weight in Ω_n attached to that n-tuple. The distinguished measurable functions[10] that tell the

[10]The measurable functions are random variables in this case since $S \subseteq \mathbb{R}$.

result of an experiment are the n coordinate functions that associate to an n-tuple ω its various entries. What is wanted is a single measure space $(\Omega, \mathcal{A}, P)$ that incorporates all these approximations. It is fairly clear that Ω should be the set of all infinite sequences of members of S and that the distinguished measurable functions are to be the infinite set of coordinate functions. Defining $\mathcal{A}$ and P is a little harder. Each n-tuple $\omega^{(n)}$ forms a singleton set in $\mathcal{A}_n$, and we associate to $\omega^{(n)}$ the set $T_n(\omega^{(n)})$ of all members of Ω whose initial segment of length n is $\omega^{(n)}$. The members of $\mathcal{A}_n$ are unions of these singleton sets, and we associate to any member X of $\mathcal{A}_n$ the union $T_n(X)$ of the sets $T_n(\omega^{(n)})$ for $\omega^{(n)}$ in X. Also, we define $P(T_n(X)) = P_n(X)$. In this way we identify $\mathcal{A}_n$ with a σ-algebra $T_n(\mathcal{A}_n)$ of subsets of Ω, and we attach a value of P to each member of $T_n(\mathcal{A}_n)$. Define

$$\mathcal{A}' = \bigcup_{n=1}^{\infty} T_n(\mathcal{A}_n).$$

The σ-algebras $T_n(\mathcal{A}_n)$ increase with n, and it follows that the union of two members of $\mathcal{A}'$ is in $\mathcal{A}'$ and that the complement of a member of $\mathcal{A}'$ is in $\mathcal{A}'$; hence $\mathcal{A}'$ is an algebra, and $\mathcal{A}$ can be taken as the smallest σ-algebra containing $\mathcal{A}'$. In the union defining $\mathcal{A}'$, a set can arise from more than one term. For example, if a set X in $\mathcal{A}_n$ is given and a set Y in $\mathcal{A}_{n+1}$ consists of all $(n+1)$-tuples whose initial n-tuple lies in X, then $T_n(X) = T_{n+1}(Y)$. The above consistency condition implies that $P_n(X) = P_{n+1}(Y)$, and hence the two definitions of P on the set $T_n(X) = T_{n+1}(Y)$ are consistent. The result is that P is well defined on $\mathcal{A}'$. Since the $T_n(\mathcal{A}_n)$ increase with n and since the restriction of P to each one is additive, it follows that P is additive. However, it is not apparent whether P is completely additive since the members of a countable disjoint sequence of sets in $\mathcal{A}'$ might not lie in a single $T_n(\mathcal{A}_n)$. This is the matter addressed by the Kolmogorov theorem.

For purposes of being able to have a general theorem, let us make an observation. Although the consistency condition used in the above example appears to rely on the ordering of the time variable, that ordering really plays no role in the above construction. We could as well have defined an F^{th} finite approximation for each finite subset F of the positive integers; the above consistency condition used in passing from $F = \{1, \ldots, n\}$ to $F' = \{1, \ldots, n, n+1\}$ implies a consistency for general finite sets of indices with $F \subseteq F'$: the result of summing the weights of all members of $\Omega_{F'}$ whose restriction to the coordinates indexed by F is a particular member of Ω_F yields the weight of the member of Ω_F. This observation makes it possible to formulate the Kolmogorov theorem in a way that allows for continuous time.

Let us then come to the example with continuous time. The example is a model of **Brownian motion**, which was discovered as a physical phenomenon in 1826. Microscopic particles, when left alone in a liquid, can be seen to move along

erratic paths; this movement results from collisions between such a particle and molecules of the liquid. An experiment can consist of a record of the position in $\mathbb{R}^3$ of a particle as a function of time. When the data are studied and suitably extrapolated to the situation that the liquid is all of $\mathbb{R}^3$, one finds an explicit formula usable to define the probability that the moving particle lies in given subsets of $\mathbb{R}^3$ at a given finite set of times. Namely, for $t > 0$, define

$$p^t(x, dy) = \frac{1}{(4\pi t)^{3/2}}\, e^{-|x-y|^2/(4t)}\, dy.$$

If $0 = t_0 < t_1 < t_2 \cdots < t_n$, if $A_0, \ldots, A_n$ are Borel sets in $\mathbb{R}^3$, and if the starting probability distribution of the particle at time 0 is a measure μ on $\mathbb{R}^3$, then the probability that the particle is in A_0 at time 0, is in A_1 at time $t_1, \ldots$, is in A_{n-1} at time t_{n-1}, and is in A_n at time t_n is to be taken as

$$\int_{x_0 \in A_0} \int_{x_1 \in A_1} \cdots \int_{x_{n-1} \in A_{n-1}} \int_{x_n \in A_n} p^{\Delta t_n}(x_{n-1}, dx_n)\, p^{\Delta t_{n-1}}(x_{n-2}, dx_{n-1}) \times \cdots \times p^{\Delta t_1}(x_0, dx_1)\, d\mu(x_0),$$

where $\Delta t_j = t_j - t_{j-1}$ for $1 \le j \le n$. Let F be $\{0, t_1, \ldots, t_n\}$. A model describing Brownian motion at the times of F takes Ω_F to be the set of functions from F into $\mathbb{R}^3$, i.e., a copy of $(\mathbb{R}^3)^{n+1}$, and the measurable sets are the Borel sets. The distinguished measurable functions are again coordinate functions;[11] they pick off the values in $\mathbb{R}^3$ at each of the times in F. Finally the measure P_F takes the value given by the above formula on the product set $A_0 \times \cdots \times A_n$, and it is evident that P_F extends uniquely to a Borel measure on $\mathbb{R}^{3(n+1)}$, the value of $P_F(A)$ for $A \subseteq \mathbb{R}^{n+1}$ being the integral over A of the integrand in the display above. If F' is the union of F and one additional time, then $P_{F'}$ and P_F satisfy a consistency property saying that if x_j is integrated over all of $\mathbb{R}^3$, then the integral can be computed and the result is the same as if index j were completely dropped in the formula; this comes down to the identity

$$\int_{y \in \mathbb{R}^3} \frac{1}{(4\pi s)^{3/2}} \frac{1}{(4\pi t)^{3/2}}\, e^{-|y-z|^2/(4s)} e^{-|x-y|^2/(4t)}\, dy = \frac{e^{-|x-z|^2/(4(s+t))}}{(4\pi(s+t))^{3/2}},$$

which follows from the formula $\int_{-\infty}^{\infty} e^{-\pi x^2}\, dx = 1$, Fubini's Theorem, and some elementary changes of variables. The passage to the limit that needs to be addressed is how to get a model that incorporates all $t \ge 0$ at once. The space can be $(\mathbb{R}^3)^{[0,+\infty)}$. An algebra $\mathcal{A}'$ can be built from the σ-algebras of Borel sets

[11] Since their values are not in $\mathbb{R}$, these measurable functions are not, strictly speaking, random variables as we have defined them in Section 1.

of the Euclidean spaces $(\mathbb{R}^3)^F$, and an additive set function P can be consistently defined on $\mathcal{A}'$ so that one recovers P_F on each space $(\mathbb{R}^3)^F$. What needs to be addressed is the complete additivity of P.

A **stochastic process** is nothing more than a family $\{x_i \mid i \in I\}$ of measurable functions defined on a measure space $(\Omega, \mathcal{A}, P)$ with $P(\Omega) = 1$. The index set I is assumed nonempty, but no other assumptions are made about it. The measurable functions have values in a more general space S than $\mathbb{R}$, but we shall assume for simplicity that S is contained in a Euclidean space $\mathbb{R}^N$ and then we may take S equal to $\mathbb{R}^N$. Although stochastic processes generally are interesting only when the measurable functions are related to each other in some special way, the Kolmogorov theorem does not make use of any such special relationship. It addresses the construction of a general stochastic process out of the approximations to it that are formed from finite subsets of I.

The situation is then as follows. Let I be an arbitrary nonempty index set, let the state space S be $\mathbb{R}^N$ for some fixed integer N, and let $\Omega = S^I$ be the set of functions from I to S. We let x_i, for $i \in I$, be the coordinate function from Ω to S defined by $x_i(\omega) = \omega(i)$. For $J \subseteq I$, we let $x_J = \{x_i \mid i \in J\}$; this is a function carrying Ω to S^J.

For each nonempty finite subset F of I, the image of x_F is the Euclidean space S^F, in which the notion of a Borel set is well defined. A subset A of Ω will be said to be **measurable of type** F if A can be described by

$$A = x_F^{-1}(X) = \{\omega \in \Omega \mid x_F \in X\} \qquad \text{for some Borel set } X \subseteq S^F.$$

The collection of subsets of Ω that are measurable of type F is a σ-algebra that we denote by $\mathcal{A}_F$. If F and F' are finite subsets of I with $F \subseteq F'$ and if the Borel set X of S^F exhibits A as measurable of type F, then the Borel subset $X \times S^{F'-F}$ of $S^{F'}$ exhibits A as measurable of type F'. Consequently $\mathcal{A}_F \subseteq \mathcal{A}_{F'}$.

Let $\mathcal{A}'$ be the union of the $\mathcal{A}_F$ for all finite F. If F and G are finite subsets of I, then we have $\mathcal{A}_F \subseteq \mathcal{A}_{F\cup G}$ and $\mathcal{A}_G \subseteq \mathcal{A}_{F\cup G}$, and it follows that $\mathcal{A}'$ is closed under finite unions and complements. Hence $\mathcal{A}'$ is an algebra of subsets of Ω.

In effect the Kolmogorov theorem will assume that we have a consistent system of stochastic processes for all finite subsets of I. In other words, for each finite subset F of I, we assume that we have a measure space $(S^F, \mathcal{B}_F, P_F)$ with $\mathcal{B}_F$ as the Borel sets of the Euclidean space S^F, with $P_F(S^F) = 1$, and with the distinguished measurable functions taken as the x_i for i in F. The measures P_F are to satisfy a consistency condition as follows. To each X in $\mathcal{B}_F$, we define a subset A_X of Ω by $A_X = x_F^{-1}(X)$; this subset of Ω is measurable of type F, and we transfer the measure from $\mathcal{B}_F$ to $\mathcal{A}_F$ by defining $P_F(A_X) = P_F(X)$. The consistency condition is that there is a well-defined nonnegative additive set function P on $\mathcal{A}'$ whose restriction to each $\mathcal{A}_F$ is P_F. The content of the theorem is that we obtain a stochastic process for I itself.

Theorem 9.5 (Kolmogorov Extension Theorem). Let I be a nonempty index set, let $S = \mathbb{R}^N$, and let $\Omega = S^I$ be the set of functions from I to S. For each nonempty finite subset F of I, let $\mathcal{A}_F$ be the σ-algebra of subsets of Ω that are measurable of type F, and let $\mathcal{A}'$ be the algebra of sets given by the union of the $\mathcal{A}_F$ for all finite F. If P is a nonnegative additive set function defined on $\mathcal{A}'$ such that $P(\Omega) = 1$ and $P\big|_{\mathcal{A}_F}$ is completely additive for every finite F, then P is completely additive on $\mathcal{A}'$ and therefore extends to a measure on the smallest σ-algebra containing $\mathcal{A}'$.

PROOF. Once we have proved that P is completely additive on $\mathcal{A}'$, P extends to a measure on the smallest σ-algebra containing $\mathcal{A}'$ as a consequence of the Extension Theorem.[12] Let A_n be a decreasing sequence of sets in $\mathcal{A}'$ with $P(A_n) \geq \epsilon > 0$ for some positive ϵ. It is enough to prove that $\bigcap_{n=1}^{\infty} A_n$ is not empty.

Each member of $\mathcal{A}'$ is measurable of type F for some finite F, and we suppose that A_n is measurable of type F_n. There is no loss of generality in assuming that $F_1 \subseteq F_2 \subseteq \cdots$ since a set that is measurable of type F is measurable of type F' for any F' containing F. Let x_i, for $i \in I$, be the i^{th} coordinate function on Ω, and let $x_F = \{x_i \mid i \in F\}$ for each finite subset F of I. Just as in the definition of joint probability distribution, we define a Borel measure μ_F on the Euclidean space S^F by $\mu_F(X) = P(x_F^{-1}(X))$. This is a measure since $P\big|_{\mathcal{A}_F}$ is assumed to be completely additive.

By definition of "measurable of type F," the set A_n is of the form

$$A_n = \left\{\omega \in \Omega \,\middle|\, x_{F_n}(\omega) \in X_n\right\}$$

for some Borel subset X_n of the Euclidean space S^{F_n}. Since $P(A_n) \geq \epsilon$, the definition of μ_{F_n} makes $\mu_{F_n}(X_n) \geq \epsilon$. Since S^{F_n} is a Euclidean space, the measure μ_{F_n} is regular. Therefore there exists a compact subset K_n of X_n with $\mu_F(X_n - K_n) \leq 3^{-n}\epsilon$. Putting

$$B_n = \left\{\omega \in \Omega \,\middle|\, x_{F_n}(\omega) \in K_n\right\},$$

we see that $P(A_n - B_n) \leq 3^{-n}\epsilon$. Let

$$C_n = \bigcap_{j=1}^{n} B_n.$$

Each C_n is a subset of A_n, and the sets C_n are decreasing. We shall prove that

$$P(C_n) \geq \epsilon/2. \qquad (*)$$

[12] Theorem 5.5 of *Basic*.

The proof of $(*)$ will involve an induction: we show inductively for each k that $B_k = D_k \cup C_k$ with $P(D_k) \leq \sum_{j=1}^{k-1} 3^{-j}\epsilon$ and $P(C_k) \geq \big(1 - \sum_{j=1}^{k} 3^{-j}\big)\epsilon$. Since $1 - \sum_{j=1}^{k} 3^{-j} \geq 1 - \sum_{j=1}^{\infty} 3^{-j} = 1 - \frac{1/3}{1-1/3} = \frac{1}{2}$, this induction will prove $(*)$. The base case of the induction is $k = 1$. In this case we have $C_1 = B_1$. If we take $D_1 = \varnothing$, then we have $B_1 = D_1 \cup C_1$ and $P(D_1) \leq 0$ trivially, and we have $P(C_1) \geq (1 - \frac{1}{3})\epsilon$ by construction of B_1. The inductive hypothesis is that $B_k = D_k \cup C_k$ with $P(D_k) \leq \sum_{j=1}^{k-1} 3^{-j}\epsilon$ and $P(C_k) \geq \big(1 - \sum_{j=1}^{k} 3^{-j}\big)\epsilon$. We know that $A_k = (A_k - B_k) \cup B_k$. Since $B_{k+1} \subseteq A_{k+1} \subseteq A_k$, we can intersect B_{k+1} with this equation and then use the inductive hypothesis to obtain

$$\begin{aligned} B_{k+1} &= (B_{k+1} \cap (A_k - B_k)) \cup (B_{k+1} \cap B_k) \\ &= (B_{k+1} \cap (A_k - B_k)) \cup (B_{k+1} \cap (D_k \cup C_k)) \\ &= (B_{k+1} \cap (A_k - B_k)) \cup (B_{k+1} \cap D_k) \cup C_{k+1}. \end{aligned}$$

If we put $D_{k+1} = (B_{k+1} \cap (A_k - B_k)) \cup (B_{k+1} \cap D_k)$, then $B_{k+1} = D_{k+1} \cup C_{k+1}$ and

$$P(D_{k+1}) \leq P(A_k - B_k) + P(D_k) \leq 3^{-k}\epsilon + \sum_{j=1}^{k-1} 3^{-j}\epsilon = \sum_{j=1}^{k} 3^{-j}\epsilon.$$

The identity $A_{k+1} = (A_{k+1} - B_{k+1}) \cup B_{k+1}$ and the inequalities $P(A_{k+1}) \geq \epsilon$ and $P(A_{k+1} - B_{k+1}) \leq 3^{-k-1}\epsilon$ together imply that $P(B_{k+1}) \geq (1 - 3^{-k-1})\epsilon$. From $B_{k+1} = D_{k+1} \cup C_{k+1}$ and $P(D_{k+1}) \leq \sum_{j=1}^{k} 3^{-j}\epsilon$, we therefore conclude that $P(C_{k+1}) \geq \big(1 - \sum_{j=1}^{k+1} 3^{-j}\big)\epsilon$. This completes the induction, and $(*)$ is thereby proved.

The set C_n is in $\mathcal{A}_{F_n}$ since $F_1 \subseteq F_2 \subseteq \cdots \subseteq F_n$, and thus C_n is given by

$$C_n = \big\{\omega \in \Omega \,\big|\, x_{F_n}(\omega) \in L_n\big\}$$

for some Borel subset L_n of K_n in S^{F_n}. For $1 \leq j \leq n$, we have

$$B_j = \big\{\omega \in \Omega \,\big|\, x_{F_n}(\omega) \in K_j \times S^{F_n - F_j}\big\},$$

and the set $K_j \times S^{F_n - F_j}$ is closed in S^{F_n} for $j < n$ and compact for $j = n$. Thus $L_n = \bigcap_{j=1}^{n} (K_j \times S^{F_n - F_j})$ is a compact subset of S^{F_n}.

If $F \subseteq F'$, let us identify $S^{F'}$ with the subset $S^{F'} \times \{0\}$ of $\Omega = S^I$, so that it is meaningful to apply x_F to $S^{F'}$. Then we have $x_F x_{F'} = x_F$, and $x_{F_n}(L_p)$ makes sense for $p \geq n$.

If $p \geq q$, then we have $x_{F_p}^{-1}(L_p) = C_p \subseteq C_q = x_{F_q}^{-1}(L_q) = x_{F_p}^{-1}(L_q \times S^{F_p - F_q})$, and hence $L_p \subseteq L_q \times S^{F_p - F_q}$. Application of x_{F_q} gives $x_{F_q}(L_p) \subseteq L_q$. If $p \geq q \geq n$, then the further application of x_{F_n} gives $x_{F_n}(L_p) \subseteq x_{F_n}(L_q) \subseteq L_n$. Thus the sets $x_{F_n}(L_p)$, as p varies for $p \geq n$, form a decreasing sequence of compact sets in S^{F_n}. Since $P(C_p) \geq \epsilon/2$ by $(*)$, C_p is not empty; thus L_p is not empty and $x_{F_n}(L_p)$ is not empty. Since L_n is a compact metric space,

$$M_n = \bigcap_{p=n}^{\infty} x_{F_n}(L_p)$$

is not empty.

Let us prove that

$$x_{F_n}(M_{n+1}) = M_n. \qquad (**)$$

For $p \geq n+1$, we have $x_{F_n}(M_{n+1}) \subseteq x_{F_n}(x_{F_{n+1}}(L_p)) = x_{F_n}(L_p)$. Intersecting the right side over p gives $x_{F_n}(M_{n+1}) \subseteq M_n$. For the reverse inclusion, let m be in M_n. Then $m = x_{F_n}(\ell_p)$ with $\ell_p \in L_p$ for $p \geq n+1$. For the same ℓ_p's, define $m'_p = x_{F_{n+1}}(\ell_p)$. Then $x_{F_n}(m'_p) = x_{F_n}(x_{F_{n+1}}(\ell_p)) = x_{F_n}(\ell_p)) = m$. The element m'_p is in $x_{F_{n+1}}(L_p)$ and hence in $\bigcap_{q=n+1}^{p} x_{F_{n+1}}(L_q)$. The elements m'_p all lie in the compact set L_{p+1}, and hence they have a convergent subsequence$\{m'_{p_k}\}$. The limit m' of this subsequence is in $\bigcap_{q=n+1}^{p_k} x_{F_{n+1}}(L_q)$ for all k, and thus m' is in M_{n+1}. Since $x_{F_n}(m'_p) = m$, we have $x_{F_n}(m') = x_{F_n}(\lim_k m'_{p_k}) = \lim_k x_{F_n}(m'_{p_k}) = m$. In other words, m lies in $x_{F_n}(M_{n+1})$. This proves $(**)$.

Using $(**)$, we shall define disjoint coordinate blocks of an element ω in Ω. Pick some m_1 in M_1, use $(*)$ to find some m_2 in M_2 with $m_1 = x_{F_1}(m_2)$, use $(*)$ to find some m_3 in M_3 with $m_2 = x_{F_2}(m_3)$, and so on. Define ω so that $x_{F_1}(\omega) = m_1$ and $x_{F_n - F_{n-1}}(\omega) = m_n - m_{n-1}$ for $n \geq 2$. Define ω to be 0 in all coordinates indexed by $I - \bigcup_{n=1}^{\infty} F_n$. Then we have

$$x_{F_n}(\omega) = x_{F_1}(\omega) + \sum_{k=2}^{n} x_{F_k - F_{k-1}}(\omega) = m_1 + \sum_{k=2}^{n} (m_k - m_{k+1}) = m_n.$$

Thus $x_{F_n}(\omega)$ is exhibited as in $M_n \subseteq L_n$ for all n. Hence ω is in $\bigcap_{n=1}^{\infty} C_n$, and we have succeeded in proving that $\bigcap_{n=1}^{\infty} C_n$ is not empty. □

Corollary 9.6. Let I be a nonempty index set, and for each i in I let μ_i be a Borel measure on $\mathbb{R}$ with $\mu_i(\mathbb{R}) = 1$. Then there exists a probability space with independent random variables x_i for i in I such that x_i has probability distribution μ_i.

PROOF. In Theorem 9.5 let $S = \mathbb{R}$, and for each finite subset F of I, define $P\big|_{\mathcal{A}_F}$ to be the product measure $\prod_{i \in F} \mu_i$ on the Euclidean space $\mathbb{R}^F$. The theorem makes R^I into a probability space by exhibiting the consistent extension P of all the $P\big|_{\mathcal{A}_F}$'s as completely additive. Then the coordinate functions x_i are the required independent random variables. □

4. Strong Law of Large Numbers

Traditional laws of large numbers concern a sequence $\{x_n\}$ of identically distributed independent random variables, and we shall assume that their common mean μ exists. Define $s_n = x_1 + \cdots + x_n$ for $n \geq 1$. The conclusion is that the quantities $\frac{1}{n}\, s_n$ converge in some sense to μ, i.e., that the x_n are Cesàro summable to the constant μ. The simplest versions of the law of large numbers assume also that the common "variance" is finite. Let us back up a moment and define this notion.

The **variance** of a random variable x with mean $E(x) = \mu$ is the quantity

$$\mathrm{Var}(x) = E\big((x-\mu)^2\big) = E(x^2) - \mu^2,$$

the right-hand equality holding since

$$E\big((x-\mu)^2\big) = E(x^2) - 2\mu E(x) + \mu^2 E(1) = E(x^2) - \mu^2.$$

For any random variables the means add since mean is linear. For two *independent* random variables x and y, the variances add since we can apply Proposition 9.3, compute the quantities

$$E((x+y)^2) = E(x^2) + 2E(xy) + E(y^2) = E(x^2) + 2E(x)E(y) + E(y^2)$$

and $$(E(x) + E(y))^2 = E(x)^2 + 2E(x)E(y) + E(y)^2,$$

and subtract to obtain

$$\mathrm{Var}(x+y)) = \big(E(x^2) - E(x)^2\big) + \big(E(y^2) - E(y)^2\big) = \mathrm{Var}(x) + \mathrm{Var}(y).$$

For a constant multiple c of a random variable x, we have

$$E(cx) = cE(x) \qquad \text{and} \qquad \mathrm{Var}(cx) = c^2\mathrm{Var}(x).$$

Returning to our sequence $\{x_n\}$ of identically distributed independent random variables, we therefore have $E(s_n) = E(x_1) + \cdots + E(x_n) = n\mu$ and $\mathrm{Var}(s_n) = \mathrm{Var}(x_1) + \cdots + \mathrm{Var}(x_n) = n\sigma^2$, where σ^2 denotes the common variance of the given random variables x_k. Consequently

$$E\big(\tfrac{1}{n}\, s_n\big) = \mu \qquad \text{and} \qquad \mathrm{Var}\big(\tfrac{1}{n}\, s_n\big) = \tfrac{1}{n}\,\sigma^2.$$

If we take our probability space to be (Ω, P) and apply Chebyshev's inequality to the variance[13] of $\frac{1}{n}\, s_n$, we obtain

$$\tfrac{1}{n}\,\sigma^2 = \int_\Omega \big(\tfrac{1}{n}\, s_n - \mu\big)^2\, dP \geq \xi^2 P\big(\{|\tfrac{1}{n}\, s_n - \mu| \geq \xi\}\big).$$

Holding ξ fixed and letting n tend to infinity, we obtain the first form historically of the law of large numbers, as follows.

[13] Chebyshev's inequality appears in Section VI.10 of *Basic* and is the elementary inequality $\int_X |f|^2\, d\mu \geq \xi^2 \mu\big(\{x \mid |f(x)| \geq \xi\}\big)$ valid on any measure space for any measurable f and any real $\xi > 0$.

Theorem 9.7 (Weak Law of Large Numbers). Let $\{x_n\}$ be a sequence of identically distributed independent random variables with a common mean μ and a common finite variance. Define $s_n = x_1 + \cdots + x_n$. Then for every real $\xi > 0$,

$$\lim_{n\to\infty} P\big(\{|\tfrac{1}{n}\, s_n - \mu| \geq \xi\}\big) = 0.$$

REMARKS. In terminology that will be defined in Section 5, the statement in words is that $\frac{1}{n}\, s_n$ "converges to μ in probability." With more effort one can obtain the conclusion of the Weak Law of Large Numbers without the hypothesis of finite variance. Instead of making that direct extra effort, however, we shall deduce in Section 5 the Weak Law of Large Numbers from the Strong Law of Large Numbers below, and there will be no need to assume finite variance.

As a practical matter, the fact that $P\big(\{|\frac{1}{n}\, s_n - \mu| \geq \xi\}\big)$ tends to 0 is of comparatively little interest. Of more interest is a probability estimate for the event that $\lim \frac{1}{n}\, s_n = \mu$. This is contained in the following theorem, whose proof will occupy the remainder of this section.

Theorem 9.8 (Strong Law of Large Numbers). Let $\{x_n\}$ be a sequence of identically distributed independent random variables whose common mean μ exists. Define $s_n = x_1 + \cdots + x_n$. Then

$$\lim_{n\to\infty} \tfrac{1}{n}\, s_n = \mu \qquad \text{with probability 1.}$$

Many members of the public have heard of this theorem in some form. Misconceptions abound, however. The usual misconception is that if the average $\frac{1}{n}\, s_n(\omega)$ has gotten to be considerably larger than μ by some point n in time, then the chances become overwhelming that the average will have corrected itself fairly soon thereafter. Independence says otherwise: that the future values of the x_k's are not influenced by what has happened through time n. In fact, if a person is persuaded that it was unreasonable for the average $\frac{1}{n}\, s_n(\omega)$ to have gotten considerably larger than μ by some time n, then the person might better instead question whether the mean μ is known correctly or even whether the individual x_n's are genuinely independent. If μ has been greatly underestimated, for example, not only was it reasonable for the average $\frac{1}{n}\, s_n(\omega)$ to have gotten considerably larger than μ, but it is reasonable for it to continue to do so.

The proof of Theorem 9.8 will be preceded by three lemmas.

Lemma 9.9 (Borel–Cantelli Lemma). Let $\{A_k\}$ be a sequence of events in a probability space (Ω, P) such that $\sum_{k=1}^{\infty} P(A_k) < \infty$. Then $P\big(\bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k\big) = 0$. Hence the probability that infinitely many of the events A_k occur is 0.

PROOF. Since $\sum_{k=1}^{\infty} P(A_k)$ is convergent, we have $\limsup_n \sum_{k=n}^{\infty} P(A_k) = 0$. For every n, we have $P\big(\bigcap_{n=1}^{\infty} \bigcup_{k\ge n} A_k\big) \le P\big(\bigcup_{k\ge n} A_k\big) \le \sum_{k=n}^{\infty} P(A_k)$. The left side of the inequality is independent of n, and therefore $P\big(\bigcap_{n=1}^{\infty} \bigcup_{k\ge n} A_k\big) \le \limsup_n \sum_{k=n}^{\infty} P(A_k) = 0$. This proves the first conclusion. Since $\bigcap_{n=1}^{\infty} \bigcup_{k\ge n} A_k$ is the set of ω that lie in infinitely many of the sets A_k, the second conclusion follows. □

Lemma 9.10. Let x be a random variable on a probability space (Ω, P). Then $\sum_{k=1}^{\infty} P(\{|x| > k\}) < \infty$ if and only if the mean of $|x|$ exists.

PROOF. Proposition 6.56b of *Basic* gives

$$\int_\Omega |x|\, dP = \int_0^\infty P(\{|x(\omega)| > \xi\})\, d\xi.$$

The lemma therefore follows from the inequalities

$$\begin{aligned}\sum_{k=1}^{\infty} P(\{|x| > k\}) &= \sum_{k=0}^{\infty} P(\{|x| > k+1\}) \le \sum_{k=0}^{\infty} \int_k^{k+1} P(\{|x| > \xi\})\, d\xi \\ &= \int_0^\infty P(\{|x| > \xi\})\, d\xi \le \sum_{k=0}^{\infty} P(\{|x| > k\}).\end{aligned}$$ □

Lemma 9.11 (Kolmogorov's inequality). Let $x_1, \dots, x_n$ be independent random variables on a probability space (Ω, P), and suppose that $E(x_k) = 0$ and $E(x_k^2) < \infty$ for all k. Put $s_k = x_1 + \cdots + x_k$. Then

$$P\big(\{\omega \mid \max(|s_1|, \dots, |s_n|) > c\}\big) \le c^{-2} E(s_n^2)$$

for every real $c > 0$.

REMARKS. It is not necessary to assume that $E(x_1) = 0$. For $n = 1$, the lemma consequently reduces to Chebyshev's inequality.

PROOF. Let A_j be the event that j is the smallest index for which $|s_j| > c$. The sets A_j are disjoint, and their union is the set whose probability occurs on the left side of the displayed inequality. Combining this fact with Chebyshev's inequality gives

$$P\big(\{\omega \mid \max(|s_1|, \dots, |s_n|) > c\}\big) = \sum_{j=1}^{n} P(A_j) \le c^{-2} \sum_{j=1}^{n} E(s_j^2 I_{A_j}), \qquad (*)$$

where I_{A_j} is the indicator function of A_j. Since $s_n = s_j + (s_n - s_j)$,

$$\begin{aligned}E(s_n^2 I_{A_j}) &= E(s_j^2 I_{A_j}) + 2E((s_n - s_j) s_j I_{A_j}) + E((s_n - s_j)^2 I_{A_j}) \\ &\ge E(s_j^2 I_{A_j}) + 2E((s_n - s_j) s_j I_{A_j}).\end{aligned}$$

The random variables $s_n - s_j$ and $s_j I_{A_j}$ are independent by Proposition 9.4, and their product has mean 0 by Proposition 9.3 since $E(s_n - s_j) = \sum_{i=j+1}^{n} E(x_i) = 0$. Therefore $E(s_n^2 I_{A_j}) \geq E(s_j^2 I_{A_j})$, and $(*)$ gives

$$P\big(\{\omega \mid \max(|s_1|, \ldots, |s_n|) > c\}\big) \leq c^{-2} \sum_{j=1}^{n} E(s_j^2 I_{A_j}) \leq c^{-2} \sum_{j=1}^{n} E(s_n^2 I_{A_j})$$
$$= c^{-2} E(s_n^2 I_{\bigcup_j A_j}) \leq c^{-2} E(s_n)^2. \qquad \square$$

PROOF OF THEOREM 9.8. Let the underlying probability space be denoted by (Ω, P). Subtraction of the constant μ from each of the random variables x_k does not affect the independence, according to Proposition 9.4, and it reduces the proof to the case that $\mu = 0$. Therefore we may proceed under the assumption that $\mu = 0$. For integers $k \geq 1$, define

$$x_k' = \begin{cases} x_k & \text{where } |x_k| \leq k, \\ 0 & \text{where } |x_k| > k, \end{cases}$$

and

$$x_k'' = \begin{cases} 0 & \text{where } |x_k| \leq k, \\ x_k & \text{where } |x_k| > k, \end{cases}$$

so that $x_k = x_k' + x_k''$. Define $s_n' = x_1' + \cdots + x_n'$ and $s_n'' = x_1'' + \cdots + x_n''$. It is enough to show that $\frac{1}{n} s_n'$ and $\frac{1}{n} s_n''$ both tend to 0 with probability 1.

First we show that $\frac{1}{n} s_n''$ tends to 0 with probability 1. Let x be a random variable with the same probability distribution as the x_k's. Referring to the definition of x_k'', we see that $P(\{|x| > k\}) = P(\{|x_k| > k\}) = P(\{x_k'' \neq 0\})$. Since $E(|x|)$ exists by assumption, Lemma 9.10 shows that $\sum_{k=1}^{\infty} P(\{|x| > k\}) < \infty$. Therefore $\sum_{k=1}^{\infty} P(\{x_k'' \neq 0\}) < \infty$. By the Borel–Cantelli Lemma (Lemma 9.10), the probability that ω lies in infinitely many of the sets $\{x_k'' \neq 0\}$ is 0. Thus by disregarding ω's in a set of probability 0, we may assume $x_k''(\omega) \neq 0$ for only finitely many k. Then $s_n''(\omega)$ remains constant as a function of n for large n, and we must have $\lim_n \frac{1}{n} s_n''(\omega) = 0$.

Now we consider $\frac{1}{n} s_n'$. The random variables x_k' are independent, but they are no longer identically distributed and they no longer need have mean 0. However, they satisfy inequalities of the form $|x_k'| \leq k$, and these in turn imply that each $E(x_k'^2)$ is finite. Concerning the means, let x be a random variable with the same probability distribution as any of the x_k's. The random variable $x_k^{\#}$ equal to x where $|x| \leq k$ and equal to 0 otherwise has $|x_k^{\#}| \leq |x|$ for all k, and hence

dominated convergence yields $\lim_k E(x_k^{\#}) = E(x) = 0$. Since x_k' and $x_k^{\#}$ have the same probability distribution, we have $\lim_k E(x_k') = 0$. The expression $E(\frac{1}{n}s_n')$ is a Cesàro sum of the sequence $\{E(x_k')\}$. Since the Cesàro sums tend to 0 when the sequence itself tends to 0, we conclude that

$$\lim_n E(\tfrac{1}{n}\, s_n') = 0. \tag{$*$}$$

Let μ be the common probability distribution of the $|x_k|$'s. The next step is to show that

$$\sum_{r=1}^{\infty} 2^{-2r} \sum_{k=2^{r-1}}^{2^r-1} E(x_k'^2) \le 2\int_0^{\infty} t\, d\mu(t). \tag{$**$}$$

The quantity on the right is twice the common value of $E(|x_k|)$ and is finite since we have assumed that the common mean of the x_k's exists. Once we have proved $(**)$, we can therefore conclude that the quantity on the left side is finite. To prove $(**)$, we write

$$\begin{aligned}\sum_{r=1}^{\infty} 2^{-2r} \sum_{k=2^{r-1}}^{2^r-1} E(x_k'^2) &= \sum_{r=1}^{\infty} 2^{-2r} \sum_{k=2^{r-1}}^{2^r-1} \int_0^k t^2\, d\mu(t) \\ &\le \sum_{r=1}^{\infty} 2^{-r} \int_0^{2^r} t^2\, d\mu(t) \\ &\le \int_0^1 t^2\, d\mu(t) + \sum_{r=1}^{\infty} 2^{-r} \int_1^{2^r} t^2\, d\mu(t).\end{aligned}$$

Let us write I and II for the two terms on the right side. The estimate for II is

$$\begin{aligned}\mathrm{II} &= \sum_{r=1}^{\infty} 2^{-r} \sum_{j=1}^{r} \int_{2^{j-1}}^{2^j} t^2\, d\mu(t) \le \sum_{r=1}^{\infty} \sum_{j=1}^{r} 2^{-r} 2^j \int_{2^{j-1}}^{2^j} t\, d\mu(t) \\ &= \sum_{j=1}^{\infty} \sum_{r=j}^{\infty} 2^{-r} 2^j \int_{2^{j-1}}^{2^j} t\, d\mu(t) = 2\sum_{j=1}^{\infty} \int_{2^{j-1}}^{2^j} t\, d\mu(t) = 2\int_1^{\infty} t\, d\mu(t).\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{I} + \mathrm{II} &\le \int_0^1 t^2\, d\mu(t) + 2\int_1^{\infty} t\, d\mu(t) \\ &\le 2\int_0^1 t\, d\mu(t) + 2\int_1^{\infty} t\, d\mu(t) = 2\int_0^{\infty} t\, d\mu(t),\end{aligned}$$

and $(**)$ is proved.

Form the sequence of random variables $x_k^* = x_k' - E(x_k')$, and put $s_n^* = x_1^* + \cdots + x_n^*$. The x_k^* are independent but no longer identically distributed. They have mean 0. Since

$$E(x_k^{*2}) = E\big((x_k' - E(x_k'))^2\big) = E(x_k'^2) - E(x_k')^2 \le E(x_k'^2),$$

$(**)$ shows that the x_k^* have the property that $\sum_{r=1}^{\infty} 2^{-2r} \sum_{k=2^{r-1}}^{2^r-1} E(x_k^{*2}) < \infty$. To prove the theorem, it would be enough to prove that the Cesàro sums $\frac{1}{n} s_n^* = \frac{1}{n} s_n' - E(\frac{1}{n} s_n')$ tend to 0, since we know from $(*)$ that $\lim_n E(\frac{1}{n} s_n') = 0$.

Changing notation, we see that we have reduced matters to proving the following: if $\{x_k\}$ is a sequence of independent random variables with mean 0 and with

$$\sum_{r=1}^{\infty} 2^{-2r} \sum_{k=2^{r-1}}^{2^r-1} E(x_k^2) < \infty, \tag{$\dagger$}$$

and if s_n denotes $x_1 + \cdots + x_n$, then $\lim_n \frac{1}{n} s_n = 0$ with probability 1.

To prove this assertion, we apply Kolmogorov's inequality (Lemma 9.11) for each $r \geq 0$ to the 2^{r-1} random variables $x_{2^{r-1}}, x_{2^{r-1}+1}, \ldots, x_{2^r-1}$. These are independent with mean 0, and $E(x_k^2)$ is finite for each by $(\dagger)$. Their partial sums are

$$s_{2^{r-1}} - s_{2^{r-1}-1}, \ldots, s_{2^r-1} - s_{2^{r-1}-1},$$

and the last partial sum has $E\big((s_{2^r-1} - s_{2^{r-1}-1})^2\big) = \sum_{k=2^{r-1}}^{2^r-1} E(x_k^2)$ by Proposition 9.3. Kolmogorov's inequality therefore gives, for any fixed $\varepsilon > 0$,

$$P\big(\{\max(|s_{2^{r-1}} - s_{2^{r-1}-1}|, \ldots, |s_{2^r-1} - s_{2^{r-1}-1}|) > 2^r\varepsilon\}\big) \leq \varepsilon^{-2} 2^{-2r} \sum_{k=2^{r-1}}^{2^r-1} E(x_k^2).$$

Summing on r and applying $(\dagger)$, we see that

$$\sum_{r=1}^{\infty} P\big(\{\max(2^{-r}|s_{2^{r-1}} - s_{2^{r-1}-1}|, \ldots, 2^{-r}|s_{2^r-1} - s_{2^{r-1}-1}|) > \varepsilon\}\big) < \infty.$$

The Borel–Cantelli Lemma (Lemma 9.9) shows that with probability 1, there are only finitely many r's for which

$$\max(2^{-r}|s_{2^{r-1}} - s_{2^{r-1}-1}|, \ldots, 2^{-r}|s_{2^r-1} - s_{2^{r-1}-1}|) > \varepsilon.$$

Fix any ω that is not in the exceptional set A_ε of probability 0, and choose $r_0 = r_0(\omega)$ such that

$$\max(2^{-r}|s_{2^{r-1}}(\omega) - s_{2^{r-1}-1}(\omega)|, \ldots, 2^{-r}|s_{2^r-1}(\omega) - s_{2^{r-1}-1}(\omega)|) \leq \varepsilon$$

for all $r \geq r_0$. If $n > 2^{r_0}$ is given, find r such that $2^{r-1} \leq n \leq 2^r - 1$. Then we have

$$\begin{aligned} 2^{-r}|s_n(\omega) - s_{2^{r-1}-1}(\omega)| &\leq \varepsilon, \\ 2^{-(r-1)}|s_{2^{r-1}-1}(\omega) - s_{2^{r-2}-1}(\omega)| &\leq \varepsilon, \\ &\vdots \\ 2^{-r_0}|s_{2^{r_0}-1}(\omega) - s_{2^{r_0-1}-1}(\omega)| &\leq \varepsilon. \end{aligned}$$

Multiplying the k^{th} inequality by 2^{-k+2}, summing for $k \geq 1$, and applying the triangle inequality, we obtain

$$n^{-1}|s_n(\omega) - s_{2^{r_0-1}-1}(\omega)| \leq 2^{-r+1}|s_n(\omega) - s_{2^{r_0-1}-1}(\omega)| \leq 4\varepsilon.$$

Therefore $$n^{-1}|s_n(\omega)| \leq 4\varepsilon + n^{-1}|s_{2^{r_0-1}-1}(\omega)|.$$

Hence $$\limsup_n \tfrac{1}{n}\,|s_n(\omega)| \leq 4\varepsilon.$$

If ω is not in the union $\bigcup_{m=1}^{\infty} A_{1/m}$ of the exceptional sets, then $\limsup_n \frac{1}{n}\,|s_n(\omega)| = 0$. This countable union of exceptional sets of probability 0 has probability 0, and the proof is therefore complete. □

5. Convergence in Distribution

The two laws of large numbers concern convergence of a sequence of random variables in two different fashions, and in this section we shall be a little more systematic about different kinds of convergence.

Let $\{x_n\}$ be a sequence of random variables on a probability space (Ω, P), and let x be another random variable on that space. One says that $\{x_n\}$ **converges almost surely** to x if $\lim_n x_n(\omega) = x(\omega)$ pointwise except possibly on a set of P measure 0. This is the notion of convergence in the Strong Law of Large Numbers (Theorem 9.8). It is same notion as almost everywhere convergence, but probabilists use a term for it that conveys something in probabilistic terms. Another expression that is used for the same notion is that $\{x_n\}$ **converges** to x **with probability** 1. Notation that is often used for this notion, but which we shall not use, is

$$\{x_n\} \xrightarrow{\text{a.s.}} x.$$

A second notion is that $\{x_n\}$ **converges in probability** to x if for each real number $\xi > 0$, $\lim_n P\big(\{\omega \,\big|\, |x_n(\omega) - x(\omega)| \geq \xi\}\big) = 0$. Some authors write

$$\{x_n\} \xrightarrow{\mathcal{P}} x.$$

This is the notion of convergence in the Weak Law of Large Numbers (Theorem 9.7). We mentioned in Section 4 that the strong law implies the weak law and that the assumption of finite variance is unnecessary in the weak law. This fact is a special case of the following result.

Proposition 9.12. If a sequence $\{x_n\}$ of random variables on a probability space (Ω, P) converges to a random variable x almost surely, then it converges to x in probability.

PROOF. Let Z be the measure-zero subset of Ω on which pointwise convergence of $\{x_n(\omega)\}$ to $x(\omega)$ fails. Redefining x and all x_n to be 0 on Z, we affect neither the hypothesis nor the conclusion, but we now have convergence at every point. Let $\xi > 0$ be given, and define

$$E_N = \{\omega \in \Omega \,\big|\, |x_n(\omega) - x(\omega)| < \xi \text{ for all } n \geq N\}.$$

Then $\{E_N\}$ is an increasing sequence of sets, and the pointwise convergence of $\{x_n\}$ to x implies that $\bigcup_N E_N = \Omega$. By the complete additivity of P, $\lim_N P(E_N) = P(\Omega) = 1$. Consequently

$$\lim_N P\{\omega \in \Omega \,\big|\, |x_n(\omega) - x(\omega)| \geq \xi \text{ for some } n \geq N\} = 0,$$

and it follows that

$$\lim_N P\{\omega \in \Omega \,\big|\, |x_N(\omega) - x(\omega)| \geq \xi\} = 0. \qquad \square$$

EXAMPLE. The expected converse statement is false. That is, it is possible for a sequence $\{x_n\}$ of random variables to converge to 0 in probability without converging to 0 almost surely. Take the probability space to be $[0, 1]$ with Lebesgue measure m, and let x_n be the indicator function of a set E_n to be specified. Then $\{x_n\}$ converges to 0 in probability if (and only if) $\lim_n m(E_n) = 0$, but it does not converge to 0 almost surely if there is a set of points ω of positive Lebesgue measure such that ω is in infinitely many of the sets E_n. To define such sets E_n, take a divergent infinite series $\sum a_n$ whose terms are positive and tend to 0, such as with $a_n = 1/n$. Let E_n be the interval extending from $\sum_{k=1}^{n-1} a_k$ to $\sum_{k=1}^{n} a_k$ but taken modulo 1. Then the sets E_n have the required properties.

There is a third kind of convergence that will interest us, and this is the kind that will occur in the Central Limit Theorem in Section 9. Let $\{x_n\}$ be a sequence of random variables, and let x be another random variable. Let F_n be the cumulative distribution function of x_n, and let F be the cumulative distribution function of x. One says that $\{x_n\}$ **converges** to x **in distribution** if $\lim_n F_n(t) = F(t)$ at every point of continuity of F. The term **converges in law** is also used, and some authors write

$$\{x_n\} \xrightarrow{\mathcal{L}} x.$$

A little surprisingly this kind of convergence is even weaker than convergence in probability. In fact, convergence in distribution depends only on the cumulative distribution functions in question. If, for example, we have any sequence

of random variables with a common distribution function, then that sequence converges in distribution. Such random variables do not even need to be defined on the same space.

To have convergence in probability, we need the the differences $|x_n(\omega) - x(\omega)|$ that appear in the definition of convergence in probability to be defined; thus x and the x_n need to be defined on the same space. So convergence in distribution does not imply convergence in probability.

Proposition 9.13. Convergence of a sequence $\{x_n\}$ of random variables in probability to a random variable x implies convergence of $\{x_n\}$ to x in distribution.

PROOF. Fix a point t where F is continuous, and fix a number $\epsilon > 0$. Since $x(\omega) > t + \epsilon$ and $|x_n(\omega) - x(\omega)| \leq \epsilon$ together imply that $x_n > t$, we have

$$\{\omega \mid x_n(\omega) \leq t\} \subseteq \{\omega \mid x(\omega) \leq t + \epsilon\} \cup \{\omega \,\big|\, |x_n(\omega) - x(\omega)| \geq \epsilon\}.$$

Hence

$$\begin{aligned} F_n(t) &= P(\{\omega \mid x_n(\omega) \leq t\}) \\ &\leq P(\{\omega \mid x(\omega) \leq t + \epsilon\}) + P(\{\omega \,\big|\, |x_n(\omega) - x(\omega)| \geq \epsilon\}) \\ &= F(t + \epsilon) + P(\{\omega \,\big|\, |x_n(\omega) - x(\omega)| \geq \epsilon\}). \end{aligned}$$

Forming the lim sup on n of this inequality and taking into account the convergence in probability of $\{x_n\}$ to x gives

$$\limsup_n F_n(t) \leq F(t + \epsilon). \tag{$*$}$$

Similarly we have

$$\{\omega \mid x(\omega) \leq t - \epsilon\} \subseteq \{\omega \mid x_n(\omega) \leq t\} \cup \{\omega \,\big|\, |x_n(\omega) - x(\omega)| \geq \epsilon\}.$$

Hence

$$F(t - \epsilon) \leq F_n(t) + P(\{\omega \,\big|\, |x_n(\omega) - x(\omega)| \geq \epsilon\}).$$

Forming the lim inf on n of this inequality and taking into account the convergence in probability of $\{x_n\}$ to x gives

$$F(t - \epsilon) \leq \liminf_n F_n(t).$$

Putting this inequality together with $(*)$, we conclude that

$$F(t - \epsilon) \leq \liminf_n F_n(t) \leq \limsup_n F_n(t) \leq F(t + \epsilon).$$

Letting ϵ tend to 0 and taking into account the continuity of F at t, we see that $\lim_n F_n(t)$ exists and equals $F(t)$. □

6. Portmanteau Lemma

It is sometimes inconvenient to use the definition of convergence in distribution to work with the notion. Fortunately some equivalent formulations are available. Some of these are identified in the following lemma. Fix a probability space (Ω, P).

Lemma 9.14 (Portmanteau[14] Lemma). Let $\{x_n\}$ be a sequence of random variables on (Ω, P), and for each n, let F_{x_n} and μ_{x_n} be the corresponding cumulative distribution function and probability distribution of x_n on $\mathbb{R}$. Let x be another random variable on (Ω, P), and let F_x and μ_x be the corresponding cumulative distribution function and probability distribution on $\mathbb{R}$. Then the following statements are equivalent:

(a) $\{x_n\}$ converges in distribution to x, i.e., $\lim_{n\to\infty} F_{x_n}(u) = F_x(u)$ at every point u of continuity of F_x,
(b) $\lim_{n\to\infty} E(g(x_n)) = E(g(x))$ for every $g \in C_{\text{com}}(\mathbb{R})$, i.e., $\{\mu_{x_n}\}$ tends to μ_x weak-star against $C_{\text{com}}(\mathbb{R})$,
(c) $\lim_{n\to\infty} E(g(x_n)) = E(g(x))$ for every bounded complex-valued continuous function g on $\mathbb{R}$.

REMARKS. In all three statements (a) through (c), the probability enters only as an overlay of interpretation: the mathematical content concerns only monotone functions and Stieltjes measures. Namely the means in (b) and (c) are nothing more than the integrals with respect to Stieltjes measures given by $E(g(x_n)) = \int_{\mathbb{R}} g\, d\mu_{x_n}$ and $E(g(x)) = \int_{\mathbb{R}} g\, d\mu_x$. Integration on the probability space (Ω, P) can be completely avoided by repeated use of Proposition 9.1.

PROOF THAT (a) IMPLIES (b). Let g be any C^1 function in $C_{\text{com}}(\mathbb{R})$, let h be the derivative of g, and let $[a, b]$ be any finite interval of $\mathbb{R}$ containing the support of g in its interior. Integration by parts (Theorem 6.53 of *Basic*) gives

$$\int_a^b F_{x_n}(t)h(t)\,dt = g(b)F_{x_n}(b) - g(a)F_{x_n}(a) - \int_a^b g\, d\mu_{x_n} = -\int_a^b g\, d\mu_{x_n} \quad (*)$$

and similarly $\int_a^b F_x(t)h(t)\,dt = -\int_a^b g\, d\mu_x$. As n tends to infinity, (a) says that $F_{x_n}(u)$ tends to $F_x(u)$ at every point of continuity of F_x. This convergence takes place everywhere except on a countable set, necessarily a Borel set of Lebesgue

[14]The etymology of the term "portmanteau" in this context is uncertain. The French word "portemanteau" in the early sixteenth century referred to a person who carried a king's cloak, and a little later the term began to refer to a traveling case, generally with two halves to it, such as would be used in riding horseback. Those definitions by themselves make the word apt for this lemma. In English, Lewis Carroll in 1882 introduced the notion of a "portmanteau word" to refer to a single word obtained by telescoping two words into one, and it can be argued that this lemma, having two ideas that belong together, is akin to a portmanteau word.

measure 0. Also the sequence is uniformly bounded by 1. Since h is bounded and has compact support, $\lim_{n\to\infty} \int_a^b F_{x_n}(t)h(t)\,dt = \int_a^b F_x(t)h(t)\,dt$ by dominated convergence. Consequently

$$\lim_n -\int_a^b g\,d\mu_{x_n} = -\int_a^b g\,d\mu_x. \tag{$**$}$$

This proves the convergence asserted by (b) for all $g \in C^1_{\text{com}}(\mathbb{R})$.

Let a general member g_0 of $C_{\text{com}}(\mathbb{R})$ be given. Corollary 6.19 and Theorem 6.20 of *Basic* show how to convolve g_0 with the members of an approximate identity of functions in $C^\infty_{\text{com}}(\mathbb{R})$ to obtain a sequence $\{g_m\}$ of members of $C^\infty_{\text{com}}(\mathbb{R})$ such that $\lim_m g_m = g_0$ uniformly. Applying $(**)$ with $g = g_m$ and passing to the limit, we obtain $(**)$ for the general member g_0 of $C_{\text{com}}(\mathbb{R})$. □

PROOF THAT (b) IMPLIES (c). We make explicit use of the fact that μ_x and all the μ_{x_n} are probability measures. Given $\epsilon > 0$, choose a finite open interval I in $\mathbb{R}$ with $\mu_x(I) \geq 1 - \varepsilon$, and fix $h \in C_{\text{com}}(\mathbb{R})$ with values in $[0, 1]$ that is 1 on I. Then $1 - h$ has values in $[0, 1]$ and is nonzero only on $\mathbb{R} - I$. So $\int_{\mathbb{R}} (1-h)\,d\mu_x \leq \mu_x(\mathbb{R} - I) = 1 - \mu_x(I) \leq \epsilon$. From $\lim \int_{\mathbb{R}} h\,d\mu_{x_n} = \int_{\mathbb{R}} h\,d\mu_x$ and $\int_{\mathbb{R}} 1\,d\mu_{x_n} = \mu_{x_n}(\mathbb{R}) = 1 = \mu_x(\mathbb{R}) = \int_{\mathbb{R}} d\mu_x$, we obtain

$$\limsup_n \int_{\mathbb{R}} (1 - h)\,d\mu_{x_n} = \int_{\mathbb{R}} (1 - h)\,d\mu_x \leq \epsilon.$$

For any continuous function g on $\mathbb{R}$ with $0 \leq g \leq 1$, we can write $g = gh + g(1 - h)$ and obtain

$$\begin{aligned}
&\left| \int_{\mathbb{R}} g\,d\mu_{x_n} - \int_{\mathbb{R}} g\,d\mu_x \right| \\
&\qquad \leq \left| \int_{\mathbb{R}} gh\,d\mu_{x_n} - \int_{\mathbb{R}} gh\,d\mu_x \right| + \int_{\mathbb{R}} (1-h)g\,d\mu_{x_n} + \int_{\mathbb{R}} (1-h)g\,d\mu_x \\
&\qquad \leq \left| \int_{\mathbb{R}} gh\,d\mu_{x_n} - \int_{\mathbb{R}} gh\,d\mu_x \right| + \int_{\mathbb{R}} (1-h)\,d\mu_{x_n} + \int_{\mathbb{R}} (1-h)\,d\mu_x.
\end{aligned}$$

Since $\lim_n \int_{\mathbb{R}} gh\,d\mu_{x_n} = \int_{\mathbb{R}} gh\,d\mu_x$, it follows that

$$\limsup_n \left| \int_{\mathbb{R}} g\,d\mu_{x_n} - \int_{\mathbb{R}} g\,d\mu_x \right| \leq 2\epsilon.$$

Because ϵ is arbitrary, $\lim_n \int_{\mathbb{R}} g\,d\mu_{x_n} = \int_{\mathbb{R}} g\,d\mu_x$. Taking linear combinations of such functions g allows us to conclude that $\lim_n \int_{\mathbb{R}} g\,d\mu_{x_n} = \int_{\mathbb{R}} g\,d\mu_x$ for every bounded continuous complex-valued function g on $\mathbb{R}$. □

PROOF THAT (c) IMPLIES (a). Suppose that $\lim_n \int_{\mathbb{R}} g\,d\mu_{x_n} = \int_{\mathbb{R}} g\,d\mu_x$ for all bounded continuous g on $\mathbb{R}$, and let u_0 be a point of continuity of F. We shall prove that $\lim_n F_{x_n}(u_0) = F_x(u_0)$.

For each subset S of $\mathbb{R}$, let I_S denote the indicator function of S. Fix a positive integer N. For all bounded continuous g on $\mathbb{R}$ with

$$I_{(-\infty,u_0]} \le g \le I_{(-\infty,u_0+1/N)},$$

integration of the left inequality with respect to μ_{x_n} and of the right inequality with respect to μ_x yields

$$F_{x_n}(u_0) = \mu_{x_n}((-\infty, u_0]) \le \int_{\mathbb{R}} g\, d\mu_{x_n}$$

and $$\int_{\mathbb{R}} g\, d\mu_x \le \mu_x((-\infty, u_0 + 1/N]) = F_x(u_0 + 1/N).$$

Thus

$$\limsup_n F_{x_n}(u_0) \le \limsup_n \int_{\mathbb{R}} g\, d\mu_{x_n} = \int_{\mathbb{R}} g\, d\mu_x \le F_x(u_0 + 1/N).$$

Since N is arbitrary and F_x is right continuous at u_0,

$$\limsup_n F_{x_n}(u_0) \le F_x(u_0). \qquad (*)$$

Fix a positive integer M. If g is a continuous function on $\mathbb{R}$ taking values in $[0, 1]$ and satisfying

$$I_{(-\infty,u_0-1/M]} \le g \le I_{(-\infty,u_0]}$$

then integration of the left inequality with respect to μ_x and of the right inequality with respect to μ_{x_n} yields

$$F_x(u_0 - 1/M) \le \int_{\mathbb{R}} g\, d\mu_x \qquad \text{and} \qquad \int_{\mathbb{R}} g\, d\mu_{x_n} \le F_{x_n}(u_0).$$

Thus

$$F_x(u_0 - 1/M) \le \int_{\mathbb{R}} g\, d\mu_x = \lim_n \int_{\mathbb{R}} g\, d\mu_{x_n} \le \liminf_n F_{x_n}(u_0).$$

Since M is arbitrary and F is left continuous at u_0,

$$F_x(u_0) \le \liminf_n F_{x_n}(u_0).$$

In combination with $(*)$, this inequality completes the proof. □

7. Characteristic Functions

Throughout this section, (Ω, P) denotes a probability space. Let x be a random variable, and let μ_x be its probability distribution. The function on $\mathbb{R}$ given as the Fourier transform of μ_x is called[15] the **characteristic function** of x and is denoted by φ_x:

$$\varphi_x(t) = \int_{\mathbb{R}} e^{-2\pi i t u}\, d\mu_x(u).$$

The notion of the Fourier transform of a finite measure on $\mathbb{R}$ was introduced in Problem 6 of Chapter VIII of *Basic*. Among other things, that problem observed that

(a) such functions are bounded and continuous,
(b) the only measure whose Fourier transform is 0 is the 0 measure.

Let us notice also that

(c) $\varphi_x(0) = 1$,
(d) $\varphi_{ax}(t) = \varphi_x(at)$.

Conclusion (d) follows from the chain of equalities

$$\begin{aligned}\varphi_x(at) &= \textstyle\int_{\mathbb{R}} e^{-2\pi i (at) u}\, d\mu_x(u) = \int_\Omega e^{-2\pi i (at) x(\omega)}\, dP(\Omega) \\ &= \textstyle\int_\Omega e^{-2\pi i t (ax)(\omega)}\, dP(\Omega) = \int_{\mathbb{R}} e^{-2\pi i t u}\, d\mu_{ax}(u).\end{aligned}$$

Conclusion (a) will be re-proved in the course of the proof of the Proposition 9.16 below.

Characteristic functions provide a viewpoint for studying probability distributions that emphasizes aspects of the distributions that are not readily apparent from their definitions. We shall see, for example, in the Levy Continuity Theorem (Theorem 9.18 below) that convergence in distribution of random variables is mirrored conveniently in convergence of the characteristic functions of the random variables. This equivalence will be a key step in establishing the Central Limit Theorem in Section 9.

Lemma 9.15. For all real x, $\left|x^{-1}(e^{ix} - 1)\right| \leq 2$.

[15] Some other authors use the term "characteristic function" to refer to a function that is 1 on some set and 0 on the complement; we have referred to this kind of function systematically as an indicator function. Still other authors use a definition of "characteristic function" involving different constants from ours, in order to be consistent with their own particular definition of the Fourier transform of a function.

PROOF. Since $|e^{ix} - 1| = |e^{-ix} - 1|$, we may assume that $x > 0$. Also we may assume that $x \leq 1$ because $x \geq 1$ certainly implies $\left|x^{-1}(e^{ix} - 1)\right| \leq 2$. For $0 < x \leq 1$, we use Taylor's Theorem (Theorem 1.36 of it Basic) in the form

$$f(x) = f(0) + f'(0)x + \int_0^x f''(s)(x - s)\, ds$$

with $f(x) = e^{ix}$. Since $f'(0) = i$ and $|f''(s)| = 1$,

$$|e^{ix} - 1| \leq |x| + \int_0^x |x - s|\, ds = |x| + \tfrac{1}{2}|x|^2 \leq |x| + \tfrac{1}{2}|x| \leq 2|x|. \qquad \square$$

Proposition 9.16. Let x be a random variable on the probability space (Ω, P), and let φ_x be its characteristic function. Then φ_x is bounded and continuous. For each integer $n \geq 1$ such that $E(x^n)$ exists, the function φ_x has n continuous derivatives and the n^{th} derivative $\varphi_x^{(n)}$ is given by n-fold differentiation under the integral defining φ_x, namely as $(-2\pi i)^n \int_{\mathbb{R}} e^{-2\pi i t u} u^n\, d\mu_x(u)$. In this case the absolute value of the n^{th} derivative $\varphi_x^{(n)}$ is bounded by $(2\pi)^n E(|x^n|)$.

PROOF. Let us prepare for an induction by carrying out a preliminary continuity step and a preliminary differentiation step.. Suppose that G is an integrable function on $\mathbb{R}$ with respect to μ_x, and let $H(t) = \int_{\mathbb{R}} e^{-2\pi i t u} G(u)\, d\mu_x(u)$. Then

$$H(t) - H(t_0) = \int_{\mathbb{R}} \left(e^{-2\pi i (t - t_0)u} - e^{-2\pi i t_0 u}\right) e^{-2\pi i t_0 u} G(u)\, d\mu_x(u),$$

and

$$|H(t) - H(t_0)| \leq \int_{\mathbb{R}} \left|e^{-2\pi i (t - t_0)u} - e^{-2\pi i t_0 u}\right| |G(u)|\, d\mu_x(u).$$

The difference of exponentials is bounded by 2, and the rest of the integrand is a fixed integrable function. Thus as t tends to t_0, we have dominated convergence, and continuity of H at t_0 follows. This completes the preliminary continuity step.

For the preliminary differentiation step. The difference quotient leading toward the derivative of $H(t) = \int_{\mathbb{R}} e^{-2\pi i t u} G(u)\, d\mu_x(u)$ is

$$\begin{aligned} h^{-1}[H(t + h) - H(t)] &= \tfrac{1}{h}\left[\int_{\mathbb{R}} (e^{-2\pi i (t+h)u} - e^{2\pi i t u}) G(u)\, d\mu_x(u)\right] \\ &= \int_{\mathbb{R}} h^{-1}(e^{-2\pi i h u} - 1) e^{-2\pi i t u} G(u)\, d\mu_x(u). \\ &= \int_{\mathbb{R}} (2\pi i h u)^{-1}(e^{-2\pi i h u} - 1)(2\pi i u) e^{-2\pi i t u} G(u)\, d\mu_x(u). \end{aligned}$$

Since Lemma 9.15 shows that $\left|(2\pi i h u)^{-1}(e^{-2\pi i h u} - 1)\right| \leq 2$, the integrand is dominated in absolute value by the single function $4\pi |u||G(u)|$ as t tends to 0. Consequently under the additional assumption that $uG(u)$ is integrable with respect to μ_x, our difference quotient is the integral of functions parametrized by h and dominated by a fixed integrable function. The integrand tends to a limit as

h tends to 0. So again we have dominated convergence, and H is differentiable with value given by differentiation under the integral sign.

With those preparations done, we can induct, starting with $n = 0$ and $G(u) = 1$. The first conclusion gives the continuity of φ_x (as a result of the integrability of 1), and the second shows that if $|u|$ is integrable, then φ_x' is differentiable with derivative given by differentiation under the integral sign. If we assume inductively the integrability of u^{n-1}, then we obtain the continuity of $\varphi_x^{(n-1)}$ immediately. With the additional assumption of integrability of u^n, we obtain the existence of $\varphi_x^{(n)}$ and the formula for computing it. The bound for the absolute value of $\varphi_x^{(n)}$ follows from the derivative formula, since $(2\pi)^n \int_{\mathbb{R}} |u^n| \, d\mu_x(u)$ equals $(2\pi)^n E(|x^n|)$. □

Proposition 9.17. If $x_1, \ldots, x_n$ are independent random variables, then their characteristic functions satisfy

$$\varphi_{x_1+\cdots+x_n} = \varphi_{x_1} \ldots \varphi_{x_n}.$$

Proof. Propositions 9.3 and 9.4 together give

$$\begin{aligned}\varphi_{x_1+\cdots+x_n}(t) &= E(e^{-2\pi i t(x_1+\cdots+x_n)}) = E(e^{-2\pi i t x_1} \cdots e^{-2\pi i t x_n}) \\ &= E(e^{-2\pi i t x_1}) \cdots E(e^{-2\pi i t x_N}) = \varphi_{x_1}(t) \cdots \varphi_{x_n}(t).\end{aligned}$$ □

8. Lévy Continuity Theorem

Let (Φ, ω) be a probability space. We shall now reformulate convergence in distribution in terms of characteristic functions.

Theorem 9.18 (Lévy Continuity Theorem). Let $\{x_n\}$ be a sequence of random variables on (Ω, P), and for each n, let F_{x_n} and φ_{x_n} be the corresponding cumulative distribution function and characteristic function of x_n on $\mathbb{R}$. Let x be another random variable on (Ω, P), and let F_x and φ_x be the corresponding cumulative distribution function and characteristic function on $\mathbb{R}$. Then the following statements are equivalent:

(a) $\{x_n\}$ converges in distribution to x, i.e., $\lim_{n\to\infty} F_n(u) = F(u)$ at every point u of continuity of F,
(b) $\{\varphi_{x_n}\}$ converges pointwise to φ_x, i.e., $\lim_{n\to\infty} \varphi_{x_n}(t) = \varphi_x(t)$ for every $t \in \mathbb{R}$.

REMARKS. In both halves of the proof, we let μ_{x_n} and μ_x be the probability distributions corresponding to F_{x_n} and F_x. We make use of the equivalence of (a) and (c) in the Portmanteau Lemma (Lemma 9.14).

PROOF THAT (a) IMPLIES (b). We apply the implication (a) implies (c) of Lemma 9.14. Since $\{x_n\}$ converges in distribution to x, the lemma shows for the function $g(u) = e^{-2\pi itu}$ that $\lim_n E(g(x_n)) = E(g(x))$, i.e., that

$$\lim_{n\to\infty} \int_{\mathbb{R}} e^{-2\pi itu}\, d\mu_{x_n}(u) = \int_{\mathbb{R}} e^{-2\pi itu}\, d\mu_x(u).$$

This is (b) of the present theorem. □

PROOF THAT (b) IMPLIES (a). Suppose that $\lim_n \varphi_{x_n}(t) = \varphi_x(t)$ pointwise for all $t \in \mathbb{R}$. According to Proposition 8.10 of *Basic*, the Fourier transform operator $\mathcal{F}$ is one-one from the Schwartz space $\mathcal{S}$ of $\mathbb{R}$ onto itself. For ψ in $C^\infty_{\text{com}}(\mathbb{R})$, $\mathcal{F}^{-1}\psi$ is therefore a well defined member of $\mathcal{S}$. In particular it is integrable. Since the functions φ_{x_n} and φ_x are bounded in absolute value by 1, dominated convergence gives

$$\lim_n \int_{\mathbb{R}} \varphi_{x_n}(t)(\mathcal{F}^{-1}\psi)(t)\, dt = \int_{\mathbb{R}} \varphi_x(t)(\mathcal{F}^{-1}\psi)(t)\, dt.$$

Substituting for the definitions in this formula, we obtain

$$\lim_n \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} e^{-2\pi itu}\, d\mu_{x_n}(u)\Big)(\mathcal{F}^{-1}\psi)(t)\, dt = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} e^{-2\pi itu}\, d\mu_x(u)\Big)(\mathcal{F}^{-1}\psi)(t)\, dt$$

On the left side of this equation, the integrand in absolute value is just $|\mathcal{F}^{-1}(\psi)(t)|$, and this is integrable for $d\mu_{x_n} \times dt$. Thus we can interchange the integrals and rewrite the left side as

$$\lim_n \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} e^{-2\pi itu}(\mathcal{F}^{-1}\psi)(t))\, dt\Big)\, d\mu_{x_n}(u) = \lim_n \int_{\mathbb{R}} \psi(u)\, d\mu_{x_n}(u).$$

Similarly we can rewrite the right side as $\int_{\mathbb{R}} \psi(u)\, d\mu_x(u)$. Thus we have

$$\lim_n \int_{\mathbb{R}} \psi(u)\, d\mu_{x_n}(u) = \int_{\mathbb{R}} \psi(u)\, d\mu_x(u) \qquad (*)$$

for every ψ in $C^\infty_{\text{com}}(\mathbb{R})$. To complete the proof, let a general member ψ_0 of $C_{\text{com}}(\mathbb{R})$ be given. We argue with an approximate identity of functions in $C^\infty_{\text{com}}(\mathbb{R})$ as at the end of the proof that (a) implies (b) in Lemma 9.14 to see that $(*)$ extends to be valid when ψ is replaced by ψ_0. Because (b) implies (a) in Lemma 9.14, the validity of $(*)$ for all ψ_0 in $C_{\text{com}}(\mathbb{R})$ completes the proof. □

9. Central Limit Theorem

We come to the main result of the chapter. Again (Ω, P) denotes a probability space.

Theorem 9.19 (Central Limit Theorem). If $\{x_n, n \geq 1\}$ is a sequence of identically distributed independent random variables on the probability space (Ω, P) with common mean μ and common nonzero finite variance σ^2 and if s_n denotes their partial sums $s_n = \sum_{k=1}^{n} x_k$, then as n tends to infinity, the random variables $\sqrt{n}\,(n^{-1}s_n - \mu)$ converge in distribution to a random variable whose cumulative distribution function has derivative $\frac{1}{\sigma\sqrt{2\pi}}\, e^{-u^2/(2\sigma^2)}$. In particular,

$$\lim_{n\to\infty} P\Big(\omega \,\Big|\, a < \sqrt{n}\left(n^{-1}s_n(\omega) - \mu\right) < b\Big) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-u^2/(2\sigma^2)}\, du$$

whenever a and b are real numbers with $a < b$.

REMARKS. The probability distribution with density $\frac{1}{\sigma\sqrt{2\pi}}\, e^{-(t-\mu)^2/2\sigma^2}\, dt$ is called the **normal distribution** with mean μ and variance σ^2. It is commonly denoted by $N(\mu, \sigma^2)$. Theorem 9.19 identifies the limiting distribution under the conditions of the theorem as $N(0, \sigma^2)$. The graph of the density function of $N(\mu, \sigma^2)$ is a familiar bell-shaped curve. Figure 9.1 shows this curve for the case that $\mu = 0$ and $\sigma = 1$.

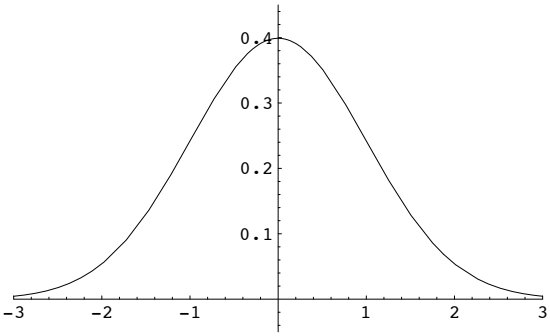


FIGURE 9.1. Graph of the density of the normal distribution $N(0, 1)$, namely the graph of $(2\pi)^{-1/2}e^{-u^2/2}$.

The situation with $\sigma^2 = 0$ is a degenerate case of the theorem, and there is still a result. In this case the random variables in question are almost surely constants, the various expressions $\sqrt{n}\,(n^{-1}s_n - \mu)$ are almost surely 0, and the limiting probability distribution is a point mass at 0.

The notation Φ is often used for the cumulative distribution function of $N(0, 1)$:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-s^2/2}\, ds,$$

and one can find extensive tables of the values of Φ. A small table of such values appears in Figure 9.2.

t	$\Phi(t) - \Phi(-t)$
0.5	.382925
1.0	.682689
1.5	.866386
2.0	.954500
2.5	.987581
3.0	.997300
3.5	.999535
4.0	.999937

FIGURE 9.2. Approximate values of $\Phi(x) - \Phi(-x)$ for the cumulative distribution function Φ of the normal distribution.

PROOF. Put $y_n = x_n - \mu$. The means of y_n and y_n^2 are

$$E(y_n) = E(x_n) - \mu = 0 \quad \text{and} \quad E(y_n^2) = E(x_n^2) - 2\mu E(x_n) + \mu^2 = E(x_n^2) - \mu^2.$$

Thus the variance of x_n, namely $\mathrm{Var}(x_n) = E(x_n^2) - \mu^2$, is equal to the variance of y_n:

$$\mathrm{Var}(y_n) = E(y_n^2) - E(y_n)^2 = E(y_n^2) = E(x_n^2) - \mu^2 = \mathrm{Var}(x_n).$$

The statement of the theorem writes σ^2 for this quantity.

We shall detect the convergence in question by the Lévy Continuity Theorem, Theorem 9.18. Let φ_y be the common characteristic function of the y_n. Since $E(y_n)$ and $E(y_n^2)$ exist, Proposition 9.16 shows that φ_y is a C^2 function. The Taylor expansion about 0 of a C^2 function φ is given for $t > 0$ by Theorem 1.36 of *Basic* as[16]

$$\begin{aligned}\varphi(t) &= \varphi(0) + \varphi'(0)t + \textstyle\int_0^t (t-s)\varphi''(s)\,ds \\ &= \varphi(0) + \varphi'(0)t + \tfrac{1}{2}\varphi''(0)t^2 + \textstyle\int_0^t \big(\varphi''(s) - \varphi''(0)\big)(t-s)^2\,ds.\end{aligned}$$

For $\varphi = \varphi_y$, Proposition 9.16 shows that $\varphi_y(0) = 1$, $\varphi_y'(0) = 0$, and

$$\varphi_y''(0) = -4\pi^2 \textstyle\int_{\mathbb{R}} u^2\,d\mu_y = -4\pi^2 E(y^2) = -4\pi^2\sigma^2.$$

Put $\alpha(t) = \int_0^t \big(\varphi''(s) - \varphi''(0)\big)(t-s)^2\,ds\big/t^2$. Since

$$\big|\textstyle\int_0^t \big(\varphi''(s) - \varphi''(0)\big)(t-s)^2\,ds\big| \le \sup_{0\le s\le t} |\varphi''(s) - \varphi''(0)|\big(\tfrac{1}{2}t^2\big),$$

[16]The case $t < 0$ is handled similarly and will be omitted.

$\alpha(t)$ is continuous for $t > 0$ and has $\lim_{t\downarrow 0}\alpha(t) = 0$. We conclude that the expansion of φ_y is

$$\varphi_y(t) = 1 - (2\pi^2\sigma^2 - \alpha(t))t^2. \tag{$*$}$$

Now we consider the characteristic functions of the random variables $\sqrt{n}\,(n^{-1}s_n - \mu)$ of the theorem. We have

$$\sqrt{n}\,(n^{-1}s_n - \mu) = n^{-1/2}\Big(\sum_{k=1}^{n} x_k - n\mu\Big) = n^{-1/2}\Big(\sum_{k=1}^{n} y_k\Big)$$

Hence

$$\begin{aligned}
\varphi_{\sqrt{n}(n^{-1}s_n-\mu)}(t) &= \varphi_{y_1+\cdots+y_n}(t/\sqrt{n}) && \text{by property (d) of characteristic functions}\\
&= \varphi_{y_1}(t/\sqrt{n})\cdots\varphi_{y_n}(t/\sqrt{n}) && \text{by independence and Proposition 9.17}\\
&= \varphi_y(t/\sqrt{n})^n && \text{by identical distributions}\\
&= \big(1-\big(2\pi^2\sigma^2-\alpha(t/\sqrt{n})\big)t^2/n\big)^n && \text{by } (*). \qquad (**)
\end{aligned}$$

Let us see that this expression has a nonzero limit as n tends to infinity, t being regarded as fixed. We shall take the logarithm of the expression, write h for $1/n$, and apply the estimate

$$|\log(1-s) + s| \le 2s^2. \tag{$\dagger$}$$

Estimate ($\dagger$) is valid for $|s| \le \frac{1}{2}$ by Taylor's Theorem (Theorem 1.36 of *Basic*) applied to the function $\log(1-s)$ about $s = 0$ and the bound of 4 on its second derivative for $|s| \le \frac{1}{2}$.

The logarithm of our expression ($**$) of interest is

$$\begin{aligned}
\log\big(1 - \big(2\pi^2\sigma^2 - \alpha(t/\sqrt{n}\,)\big)t^2/n\big)^n &= n\log\big(1 - (2\pi^2\sigma^2 - \alpha(t/\sqrt{n}\,)\big)t^2/n\big)\\
&= h^{-1}\log\big(1 - h\big(2\pi^2\sigma^2 - \alpha(t\sqrt{h}\,)\big)t^2\big)
\end{aligned}$$

with $h = 1/n$, and ($\dagger$) says that

$$\begin{aligned}
\big|\log\big(1 - h\big(2\pi^2\sigma^2 - \alpha(t\sqrt{h}\,)\big)t^2\big) + h\big(2\pi^2\sigma^2 - \alpha(t\sqrt{h}\,)\big)t^2\big|&\\
\le 2\big(h\big(2\pi^2\sigma^2 - \alpha(t\sqrt{h}\,)\big)t^2\big)^2&
\end{aligned}$$

if the side condition

$$\big|h\big(2\pi^2\sigma^2 - \alpha(t\sqrt{h}\,)\big)t^2\big| \le \tfrac{1}{2} \tag{$\dagger\dagger$}$$

is satisfied. Here t is fixed, and (††) is satisfied for positive h small enough because $\lim_{h\downarrow 0} \alpha(t\sqrt{h}\,) = 0$. Thus the logarithm of (∗∗) satisfies

$$\begin{aligned}\Big|\log\big(1 - \big(2\pi^2\sigma^2 - \alpha(t/\sqrt{n}\,)\big)t^2/n\big)^n + \big(2\pi^2\sigma^2 - \alpha(t\sqrt{h}\,)\big)t^2\Big| \\ \leq 2h\Big(\big(2\pi^2\sigma^2 - \alpha(t\sqrt{h}\,)\big)t^2\Big)^2,\end{aligned}$$

where $h = 1/n$. The right side tends to 0 as n tends to infinity and h tends to 0, and so does $\alpha(t\sqrt{h}\,)t^2$. Thus

$$\lim_{n\to\infty} \log\big(1 - \big(2\pi^2\sigma^2 - \alpha(t/\sqrt{n}\,)\big)t^2/n\big)^n = -2\pi^2\sigma^2t^2.$$

Since exp is a continuous function, we can exponentiate both sides and interchange exponential and limit, obtaining

$$\lim_{n\to\infty} \big(1 - \big(2\pi^2\sigma^2 - \alpha(t/\sqrt{n}\,)\big)t^2/n\big)^n = e^{-2\pi^2\sigma^2t^2}.$$

In other words, the characteristic functions of the random variables $\sqrt{n}(n^{-1}s_n - \mu)$ satisfy

$$\lim_{n\to\infty} \varphi_{\sqrt{n}(n^{-1}s_n-\mu)}(t) = e^{-2\pi^2\sigma^2t^2}$$

pointwise for all t.

To apply Theorem 9.18 and complete the proof, we need to identify a probability measure whose Fourier transform is $e^{-2\pi^2\sigma^2t^2}$. That is, we want the inverse Fourier transform of $e^{-2\pi^2\sigma^2t^2}$. According to Proposition 8.2 of *Basic* and the remarks afterward, the function $e^{-\pi u^2}$ has Fourier transform $e^{-\pi t^2}$, and for $a > 0$, the Fourier transform of the dilate $u \mapsto a^{-1}e^{-\pi a^{-2}u^2}$ is $t \mapsto e^{-\pi a^2t^2}$. Thus the function $\frac{1}{\sigma\sqrt{2\pi}}\, e^{-u^2/(2\sigma^2)}$ has Fourier transform $e^{-2\pi^2\sigma^2t^2}$. In other words, the sequence $\{\sqrt{n}\,(n^{-1}s_n - \mu)\}$ of random variables on (Ω, P) converges in distribution to the random variable given by the coordinate function u on the probability space $\big(\mathbb{R}, \frac{1}{\sigma\sqrt{2\pi}}\, e^{-u^2/(2\sigma^2)}\, du\big)$. □

EXAMPLES.

(1) Flipping a large number of coins results in a normal distribution for the total number of heads. This special case of the Central Limit Theorem is the Theorem of de Moivre and Laplace and is the subject of Problem 18 at the end of the chapter.

(2) Brownian motion, as discussed near the beginning of Section 3. The collisions of molecules with a microscopic particle impart small changes in the path of a particle, and the overall motion of the particle can be analyzed in a discrete model as a sum of independent random variables. Then it is natural to expect that the motion is governed by the exponential of a quadratic expression, and the formulas of the beginning of Section 3 are forced on the model. This effect comes about from the Central Limit Theorem,

(3) In the Central Limit Theorem, suppose that each of the random variables x_j has the normal distribution $N(0, \sigma^2)$. In this case we can compute exactly what is happening within the proof of the theorem. The characteristic function φ_y of $N(0, \sigma^2)$ is the Fourier transform of the function $(2\pi\sigma^2)^{1/2}e^{-u^2/(2\sigma^2)}$, which is $e^{-2\pi\sigma^2 t^2}$, as was observed near the end of the proof. In other words the function φ_y in the proof is exactly

$$\varphi_y(t) = e^{-2\pi\sigma^2 t^2}.$$

Since $\mu = 0$, we obtain

$$\varphi_{\sqrt{n}(n^{-1}s_n - \mu)}(t) = \varphi_y(t/\sqrt{n})^n,$$

just as in $(**)$ of the proof. We have an exact expression for φ_y, and thus

$$\varphi_{\sqrt{n}(n^{-1}s_n - \mu)}(t) = (e^{-2\pi\sigma^2(t/\sqrt{n})^2})^n = e^{-2\pi\sigma^2 t^2}$$

exactly. Taking the inverse Fourier transform shows that the probability distribution of $\sqrt{n}(n^{-1}s_n - \mu)$ is exactly $\frac{1}{\sigma\sqrt{2\pi}}\, e^{-u^2/(2\sigma)^2}\, du$, which is $N(0, \sigma^2)$. In other words, all the terms of the sequence are the same, and the convergence is trivial.

So far in this chapter, we have seen that probability theory establishes models that in principle can be used to generate data about future events. In real-world applications, one wants to work in the opposite direction—taking some data from past events, extracting parameters to be able to construct a probability model, and comparing the given data with what happens in that model. This is the question of statistical inference, a subject in the field of statistics. We look at one aspect of that question in the next section.

10. Statistical Inference and Gosset's t Distribution

The Central Limit Theorem is an important tool used in extracting information in real-world applications. In many practical cases one works with a (very large) population but measures some property in only some of the possible cases, those in a sample. Let us concentrate on the mean value. Typically one wants to estimate the mean value of this property for the whole population but as a practical matter can compute it only for the sample. One then wants to extrapolate and use the mean of the sample as the mean of the whole population. The difficulty is in saying how reliable this extrapolated mean is likely to be, that the answer is within such-and-such interval with a probability of at least a certain amount.[17]

[17]This interval is often called the **margin of error**. The usual convention unless an author states otherwise is that the probability of being within the margin of error is at least .95.

Examples of such cases are voter preferences for candidates for a particular election and the efficacy of a new drug in medicine. In these examples the population is made up of people, and each person in the sample is associated with a random variable x_n. The possible values of x_n and their probabilities give a probability distribution associated with that person. The Central Limit Theorem then gives an idea what to expect of the general population, *provided the hypotheses of independence and identical distributions are satisfied*.[18]

In practice certain known potential dependences need to be taken into account. With voting preferences it is known that a voter's age, gender, income, education, and political affiliation may affect the voting preference. In testing a drug, it is known that a person's age, gender, and health history may affect how well a drug works. The idea is to work with each category separately, assuming that its members are more or less independent, and to apply the Central Limit Theorem to each category. Then the results from the different categories are combined with some weighting. We shall not pursue this question of handling potential dependences but shall concentrate on situations in which the entire population is assumed to be independent and identically distributed.

W. S. Gosset studied this situation in a famous 1908 paper. He was an employee of a brewery in Ireland, and one of his interests was in assessing the chemical properties of barley on the basis of a rather small sample. The opening paragraphs of his paper[19] read as follows:

> Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.
>
> Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a greater number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.
>
> If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty: (1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation

[18] There are versions of the Central Limit Theorem that relax the assumptions of independence and identical distributions somewhat, but these versions will not be of concern to us.

[19] W. S. Gosset (writing under the pseudonym "Student"), "The probable error of a mean," *Biometrika* 6 (1), 1–25.

> so close that a small sample will give no information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed, yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.
>
> The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use tables of the probability integral.
>
> But, as we decrease the number of experiments, the value of the standard deviation found from the sample of experiments becomes itself subject to an increasing error, until judgments reached in this way become altogether misleading.
>
> In routine work . . . .
>
> There are other experiments, however, which cannot easily be repeated very often; in such cases it is sometimes necessary to judge the certainty of the results from a very small sample, which itself affords the only indication of the variability. Some chemical, many biological, and most agricultural and large-scale experiments belong to this class, which has hitherto been almost outside the range of statistical inquiry.
>
> . . . The aim of the present paper is to determine the point at which we may use the tables of the probability integral [pertaining to the Central Limit Theorem] in judging of the significance of the mean of a series of experiments, and to furnish alternative tables for use when the number of experiments is too few.

In the language we have been using, Gosset worked with a sample from a large population. His random variables $x_1, \ldots, x_n$ picked out some numerical property of each member of a sample of size n from the population. It is helpful to regard each member ω of the the underlying probability space Ω as one possible situation, as far as those n random variables are concerned. Gosset assumed that $x_1, \ldots, x_n$ were independent and identically distributed, and he assumed further that the common probability distribution of the x_j's was a normal distribution $N(\mu, \sigma^2)$ in which the mean μ and the variance σ^2 were unknown.[20]

[20] In real-life applications the common probability distribution is not likely to be exactly normal, but it is often approximately normal. In this situation the usual practice is to proceed as if the

He introduced the **sample mean**

$$\bar{x}(\omega) = n^{-1}(x_1(\omega) + \cdots + x_n(\omega))$$

and the **sample variance** defined by

$$s(\omega)^2 = \frac{1}{n-1} \sum_{j=1}^{n} \big(x_j(\omega) - \bar{x}(\omega)\big)^2,$$

which have

$$E(\bar{x}) = \mu \qquad \text{and} \qquad E(s^2) = \sigma.$$

The latter equality is what accounts for the coefficient $\frac{1}{n-1}$ in the definition of s^2. Gosset worked with a random variable

$$t(\omega) = \frac{\bar{x}(\omega) - \mu}{s(\omega)/\sqrt{n}}.$$

It has an interpretation as the difference between the locations of the sample mean and the true mean, divided by the sample standard deviation, all multiplied by the same normalizing factor $\sqrt{n}$ as in the statement of the Central Limit Theorem. In an allusion to Gosset's pseudonym Student, the probability distribution of t is often called **Student's t distribution**, but we shall follow the simpler convention of calling it **Gosset's t distribution**. It has the single parameter n, and for reasons that will emerge below, one refers to it as the t distribution "with $n-1$ **degrees of freedom**." The t distribution can therefore be used to estimate how likely the true mean is to lie in a given interval about the sample mean.

Theorem 9.20. If $n > 1$ and if $x_1, \ldots, x_n$ are independent random variables with the common probability distribution $N(\mu, \sigma^2)$, then the density $f_n(t)$ of the t distribution with $n-1$ degrees of freedom is given by[21]

$$f_n(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

The proof will be given at the end of this section. Observe that $f_n(t)$ depends neither on μ nor on σ. Because of this property, Gosset's t distribution is indeed usable for estimating how likely the true mean μ is to lie in a given interval about the sample mean $\bar{x}(\omega)$. We shall give an example in a moment.

---

distribution were exactly normal but to be aware that some errors may be introduced through the approximation. We shall not pursue this matter.

[21] The formula makes use of the gamma function, defined by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ for $x > 0$. This function was studied in Proposition 6.34 of *Basic*, which showed that $\Gamma(x+1) = x\Gamma(x)$ for $x > 0$, $\Gamma(1) = 1$, $\Gamma(n+1) = n!$ for integers $n \geq 0$, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

The coefficient c_n in the expression for $f_n(t)$ should be viewed as some harmless positive constant; the function $(1 + (n-1)^{-1}t^2)^{-n/2}$ is integrable on $\mathbb{R}$, and the coefficient makes $\int_{\mathbb{R}} f_n(t)\,dt = 1$. If we write c_n separately for n even and n odd, the value is[22]

$$c_n = \begin{cases} \dfrac{(n-2)(n-4)\cdots 5\cdot 3}{2\sqrt{n-1}\,(n-3)(n-5)\cdots 4\cdot 2} & \text{for } n \text{ odd} \\ \dfrac{(n-2)(n-4)\cdots 4\cdot 2}{\pi\sqrt{n-1}\,(n-3)n-5)\cdots 5\cdot 3} & \text{for } n \text{ even.} \end{cases}$$

Problem 20 at the end of the chapter shows that the coefficient c_n has a finite nonzero limit as n tends to infinity and that

$$\lim_{n\to\infty} f_n(t) = \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2},$$

i.e., that $f_n(t)$ converges pointwise to the density of the normal distribution $N(0, 1)$. This convergence is illustrated in Figure 9.3. The density of the t distribution is smaller in the center than that of the normal distribution, and it has larger tails. As n increases, this effect becomes less pronounced.

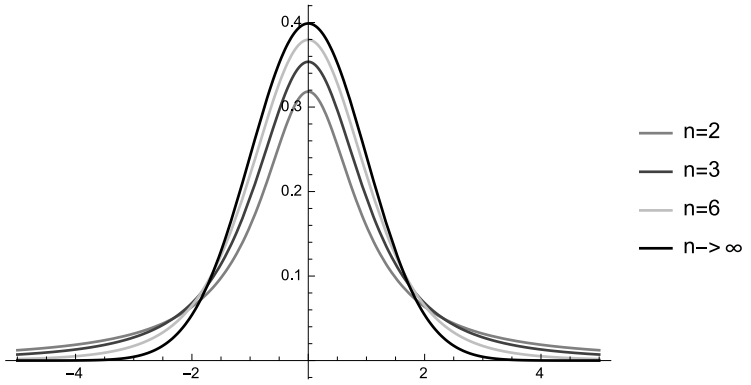


FIGURE 9.3. Normal distribution as a limit of Gosset's t distribution.

The opening of Gosset's 1908 paper mentioned his forming some tables of his distribution for various values of n. He stopped forming these tables with $n = 30$,

[22]Many authors refer to $f_n(t)$ and its coefficient c_n by using the number of degrees of freedom $\nu = n - 1$ as parameter and then using $\nu + 1$ in place of n in all tables and formulas. The role of the number of degrees of freedom will become a little clearer in the course of the proof of Theorem 9.20.

apparently regarding the t distribution with 29 degrees of freedom as close enough to the normal distribution $N(0, 1)$ to make further tables unnecessary. Nowadays with high-speed computers, one does need to draw a distinction between small values of n and large values; instead one can use the t distribution in all cases. The table in Figure 9.4 shows the minimal choice of c needed so that $\int_{-c}^{c} f_n(t)\,dt$ is $\geq .95$, $\geq .99$, and $\geq .995$.

n\Threshold	.95	.99	.995
2	12.7065	63.6551	127.3447
3	4.3026	9.9247	14.0887
5	2.7764	4.6041	5.5976
7	2.4469	3.7074	4.3168
10	2.2621	3.2498	3.6896
12	2.2010	3.1058	3.4966
15	2.1448	2.9768	3.3257
20	2.0930	2.8609	3.1737
30	2.0452	2.7564	3.0380
50	2.0096	2.6800	2.9397
$\to \infty$	1.9600	2.5759	2.8071

FIGURE 9.4. Table of approximate minimal values of c such that $\int_{-c}^{c} f_n(t)\,dt$ exceeds a threshold.[23]

EXAMPLE. A manufacturer of light bulbs claims in its advertising that one type of its bulbs last for 1000 hours. A consumer advocate randomly selects 10 bulbs for testing. The sampled bulbs last a mean of 950 hours with a standard deviation of 50 hours. If the advertising claim were true, what is the probability that 10 randomly selected bulbs would have an average life of no more than 950 hours? To answer this question, we use the given data to compute a t score. The computation gives

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{950 - 1000}{50/\sqrt{10}} = -\sqrt{10} \approx -3.16,$$

and Figure 9.4 shows that the probability is really small; more precisely the probability is approximately $\int_{-\infty}^{-3.16} f_{14}(t)\,dt \approx .0058$. The conclusion is that the advertising probably exaggerates the lifetime of the light bulbs.

[23] For a larger table see `http://www.easycalculation.com/statistics/t-distribution/t-distribution-critical-value-table.php`, from which this small table was extracted in February 2016.

We turn to the proof of Theorem 9.20, beginning with some preliminary work with some specific probability distributions.

The first of these is the **gamma distribution** with parameters $\alpha > 0$ and $\lambda > 0$. It is given by[24]

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx \qquad \text{on } (0, +\infty).$$

It is taken to be 0 for $x \le 0$. To check that this is a probability distribution, we need to check that the coefficient of dx has total integral 1. In fact, the change of variables $y = \lambda x$ gives

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1} \lambda^{1-\alpha} e^{-y} \lambda^{-1}\, dy = 1,$$

as required.

Lemma 9.21. Suppose that $x_1, \ldots, x_n$ are independent random variables and that for $1 \le j \le n$, the random variable x_j has the gamma distribution with parameters α_j and λ. Then $x_1 + \cdots + x_n$ has the gamma distribution with parameters $\alpha_1 + \cdots + \alpha_n$ and λ.

REMARK. The proof will use facts about characteristic functions from Section 7 and also elementary complex analysis as in Appendix B of *Basic*.

PROOF. Fix $\alpha > 0$. For $\operatorname{Re} z > 0$, let $f(z) = \int_0^\infty z^\alpha x^{\alpha-1} e^{-zx}\, dx$. If z is real and positive, then the change of variables $y = zx$ shows that $f(z) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy = \Gamma(\alpha)$. In other words, $f(z)$ is the constant $\Gamma(\alpha)$ for z real. Let us show that $f(z)$ is analytic, and then we can conclude that $f(z) = \Gamma(\alpha)$ for all z with $\operatorname{Re} z > 0$.

The integrand for $f(z)$ is continuous for (z, x) in the set $\{\operatorname{Re} z > 0\} \times (0, \infty)$, and it is analytic in the first variable. Lemma B.12 and Corollary B.15 of *Basic* show that $f_{\varepsilon,N}(z) = \int_\varepsilon^N z^\alpha x^{\alpha-1} e^{-zx}\, dx$ is analytic for $\operatorname{Re} z > 0$ whenever $0 < \varepsilon < N < \infty$, and as ε tends to 0 and N tends to ∞, $f_{\varepsilon,N}(z)$ converges to $f(z)$ uniformly on compact subsets of z with $\operatorname{Re} z > 0$. Consequently $f(z)$ is analytic for $\operatorname{Re} z > 0$. Since $f(z)$ is constantly equal to $\Gamma(\alpha)$ for z real and positive, $f(z)$ is constantly equal to $\Gamma(\alpha)$ for $\operatorname{Re} z > 0$. Taking $z = \lambda + 2\pi i t$ therefore gives

$$\int_0^\infty (\lambda + 2\pi i t)^\alpha x^{\alpha-1} e^{-(\lambda+2\pi i t)x}\, dx = \Gamma(\alpha) \qquad (*)$$

for all real t.

[24] *Warning:* Some authors define the gamma distribution to have parameters α and β, where β is the reciprocal of the parameter λ here.

Now we can compute the characteristic function φ_x of a random variable x having the gamma distribution with parameters α and λ. This is just the Fourier transform of the gamma distribution itself. Specifically

$$\begin{aligned}\varphi_x(t) &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} e^{-2\pi i x t}\, dx \\ &= \frac{1}{\Gamma(\alpha)} \Big(\frac{\lambda}{\lambda + 2\pi i t}\Big)^\alpha \int_0^\infty (\lambda + 2\pi i t)^\alpha x^{\alpha-1} e^{-(\lambda+2\pi i t)x}\, dx\end{aligned}$$

Applying $(*)$, we see that

$$\varphi_x(t) = \Big(\frac{\lambda}{\lambda + 2\pi i t}\Big)^\alpha. \tag{**}$$

Finally we can prove the lemma. Equation $(**)$ shows for each j that $\varphi_{x_j}(t) = \big(\frac{\lambda}{\lambda+2\pi i t}\big)^{\alpha_j}$. The assumed independence and Proposition 9.17 therefore give

$$\varphi_{x_1+\cdots+x_n}(t) = \prod_{j=1}^n \varphi_{x_j}(t) = \prod_{j=1}^n \big(\tfrac{\lambda}{\lambda+2\pi i t}\big)^{\alpha_j} = \big(\tfrac{\lambda}{\lambda+2\pi i t}\big)^{\alpha_1+\cdots+\alpha_n}.$$

By $(**)$ the right side is the Fourier transform of the gamma distribution with parameters $\alpha_1 + \cdots + \alpha_n$ and λ. Since the Fourier transform operator is one-one on $L^1(\mathbb{R})$, $x_1 + \cdots + x_n$ has the gamma distribution with parameters $\alpha_1 + \cdots + \alpha_n$ and λ. □

The second specific probability distribution that we need is the distribution of $u_1^2 + \cdots + u_k^2$ if $u_1, \ldots, u_k$ are independent random variables with the common normal distribution $N(0, 1)$. This called the **chi-square distribution with** k **degrees of freedom**. The notation $\chi^2(k)$ is used for this distribution.

Lemma 9.22. The chi-square distribution with k degrees of freedom equals the gamma distribution with parameters $\alpha = \frac{k}{2}$ and $\lambda = \frac{1}{2}$ and is given by

$$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}\, dx \qquad \text{on } (0, \infty).$$

It is 0 for $x \le 0$.

PROOF. First consider $k = 1$. The statement that u_1 has the distribution $N(0, 1)$ means that the probability distribution of u_1 is

$$(2\pi)^{-1/2} e^{-x^2/2}\, dx.$$

This implies for $a \geq 0$ and $b > 0$ that

$$P(a < u_1(\omega)^2 < b) = P(\sqrt{a} < |u_1(\omega)| < \sqrt{b}) = 2\int_{\sqrt{a}}^{\sqrt{b}} (2\pi)^{-1/2} e^{-x^2/2}\, dx.$$

We apply the change of variables $y = x^2$ and see that the above expression is

$$= 2^{1/2} \int_a^b \pi^{-1/2} e^{-y/2} \tfrac{1}{2} y^{-1/2}\, dy = 2^{-1/2}\pi^{-1/2} \int_a^b y^{-1/2} e^{-y/2}\, dy.$$

From this we can conclude that the probability distribution of u_1^2 is

$$2^{-1/2}\Gamma(\tfrac{1}{2}) y^{-1/2} e^{-y/2}\, dy \qquad \text{on } (0, +\infty),$$

which is the gamma distribution with parameters $\alpha = \frac{1}{2}$ and $\lambda = \frac{1}{2}$.

Now consider general k. The independence of $u_1, \ldots, u_k$ implies that $u_1^2, \ldots, u_k^2$ are independent, according to Proposition 9.4. Each of them has the gamma distribution with parameters $\alpha = \frac{1}{2}$ and $\lambda = \frac{1}{2}$, and we conclude from Lemma 9.21 that $u_1^k + \cdots + u_k^2$ has the gamma distribution with parameters $\alpha = \frac{k}{2}$ and $\lambda = \frac{1}{2}$. □

Lemma 9.23. If w and v are independent random variables such that w has the distribution $N(0, 1)$ and v has the distribution $\chi^2(k)$ with $k > 0$, then the random variable t defined by

$$t = \frac{w}{\sqrt{v/k}}$$

has Gosset's t distribution with k degrees of freedom, namely

$$\frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})} \Big(1 + \frac{t^2}{k}\Big)^{-\frac{k+1}{2}}$$

REMARKS. In our application of this lemma to the proof of Theorem 9.20, the integer k will be $n - 1$, not n. In the notation of Theorem 9.20, the exact distributions to which the lemma will be applied are

$$w = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\Big(\frac{\bar{x} - \mu}{\sigma}\Big)$$

and

$$v = \frac{(n-1)s^2}{\sigma^2}.$$

To apply the lemma, we shall need to prove that w has probability distribution $N(0, 1)$, v has probability distribution $\chi^2(n-1)$, and w and v are independent.

PROOF. We use the technique of Example 3b in Section 2. By independence the joint probability distribution of (w, v) is

$$\frac{1}{\sqrt{2\pi}}\, e^{-w^2/2}\, \frac{1}{2^{k/2}\Gamma(k/2)}\, v^{k/2-1} e^{-v/2}\, dw\, dv \qquad \text{on } \mathbb{R} \times (0, +\infty).$$

Define a new random variable t by $t = w/\sqrt{v/k}$, and change variables from (w, v) to (t, u) using the transformation with $t = w/\sqrt{v/k}$ and $u = v$. Here (t, u) lies in $\mathbb{R} \times (0, +\infty)$, and $(w, v) \mapsto (t, u)$ is one-one onto. The inverse transformation has $w = t\sqrt{u/k}$ and $v = u$, and the Jacobian matrix of the inverse transformation is

$$\begin{pmatrix} \partial w/\partial t & \partial w/\partial u \\ \partial v/\partial t & \partial v/\partial u \end{pmatrix} = \begin{pmatrix} \sqrt{u/k} & \frac{1}{2}t/\sqrt{ku} \\ 0 & 1 \end{pmatrix}.$$

Thus $dw\, dv = \sqrt{u/k}\, dt\, du$, and the joint probability distribution in the new variables is

$$\frac{\sqrt{u/k}}{\sqrt{2\pi}\, 2^{k/2}\Gamma(k/2)}\, e^{-t^2u/(2k)} u^{k/2-1} e^{-u/2}\, dt\, du.$$

To obtain the probability distribution of w, we integrate out the variable u for $u \in \mathbb{R}$. We need to compute

$$\frac{1/\sqrt{k}}{\sqrt{\pi}\,\Gamma(k/2)} \int_{-\infty}^{\infty} e^{-u(1+t^2/k)/2} (u/2)^{(k+1)/2} u^{-1}\, du. \qquad (*)$$

We use the change of variables $x = u(1 + t^2/k)/2$ to see that $(*)$ is

$$\begin{aligned} &= \frac{1}{\sqrt{k\pi}\,\Gamma(k/2)} \int_{-\infty}^{\infty} e^{-x} x^{(k+1)/2} (1 + t^2/k)^{-(k+1)/2} x^{-1}\, dx \\ &= \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\,\Gamma(k/2)}\, (1 + t^2/k)^{-(k+1)/2}, \end{aligned}$$

as required. □

An n-by-n real matrix A is said to be **orthogonal** if it satisfies $AA^{\mathrm{tr}} = 1$. In this case we have also $A^{\mathrm{tr}}A = 1$. The condition $AA^{\mathrm{tr}} = A^{\mathrm{tr}}A = 1$ is equivalent to the condition that the columns of A are orthogonal under the dot product and have length 1, and similarly for rows. The linear transformation corresponding to an orthogonal matrix preserves volumes and therefore has determinant ± 1.

Lemma 9.24. Let $p_1, \ldots, p_n$ be independent random variables with $N(0, 1)$ as probability distribution, and let A be an n-by-n orthogonal matrix. Then the random variables $y_1, \ldots, y_n$ defined by $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = A \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$ are independent and have $N(0, 1)$ as probability distribution.

REMARK. This lemma is an illustration of the principle in Example 4 in Section 2.

PROOF. Write p and y for the respective column vectors $p = (p_1, \dots, p_n)$ and $(y_1, \dots, y_n)$. The probability distribution of each p_j is $(2\pi)^{-1/2} e^{-p_j^2/2}\, dp_j$ by assumption. In terms of the dot product, the independence implies that the joint probability distribution of $p_1, \dots, p_n$ is

$$(2\pi)^{-n/2} \prod_{j=1}^{n} e^{-(p_1^2+\cdots+p_n^2)/2}\, dp_1 \cdots dp_n = (2\pi)^{-n/2} e^{-(p\cdot p)/2}\, dp_1 \cdots dp_n. \quad (*)$$

Since A is an orthogonal matrix, $y \cdot y = Ap \cdot Ap = p \cdot A^{\mathrm{tr}} Ap = p \cdot p$, and also $|\det A| = 1$. Thus under the change of variables $y = Ap$, the expression $(*)$ is

$$= (2\pi)^{-n/2} e^{-(y\cdot y)/2}\, dy_1 \cdots dy_n = (2\pi)^{-n/2} \prod_{j=1}^{n} e^{-(y_1^2+\cdots+y_n^2)/2}\, dy_1 \cdots dy_n$$

$$= \prod_{j=1}^{n} (2\pi)^{-1} e^{-y_j^2/2}\, dy_j$$

The fact that the joint probability distribution splits as a product proves the independence of $y_1, \dots, y_n$, and we can read off from this formula that the probability distribution of y_j is $(2\pi)^{-1} e^{-y_j^2/2}\, dy_j$, i.e., that y_j has probability distribution $N(0, 1)$. □

PROOF OF THEOREM 9.20. We define random variables $w = \sigma^{-1}\sqrt{n}(\bar{x} - \mu)$ and $v = \sigma^{-2}(n-1)s^2$ as in the remarks with Lemma 9.23. Then $v/k = v/(n-1) = \sigma^{-2}s^2$, and $\sqrt{v/k} = \sigma^{-1}s$. The quotient in the statement of Lemma 9.23, which is labeled as t, becomes

$$\frac{w}{\sqrt{v/k}} = \frac{\sigma^{-1}\sqrt{n}\,(\bar{x} - \mu)}{\sigma^{-1}s} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

and matches the random variable t in the statement of Theorem 9.20. Theorem 9.20 will therefore follow from Lemma 9.23 if we show that w has probability distribution $N(0, 1)$, v has probability distribution $\chi^2(n-1)$, and w and v are independent.

Let A be any n-by-n orthogonal matrix whose first row has every entry equal to $1/\sqrt{n}$. For example, we can start from

$$B = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ -\frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots & 0 \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & 0 & \cdots & 0 \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & \cdots & 0 \\ & \vdots & & & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{pmatrix}.$$

Observe that the rows of B are orthogonal under dot product. Define a matrix A to be the same as B except that the j^{th} row, for $j \geq 2$, is to be multiplied by $\sqrt{j/(j-1)}$. Then the rows of A are orthogonal and have length 1, so that A is an orthogonal matrix.[25]

Define random variables p_j for $1 \leq j \leq n$ by $p_j = \sigma^{-1}(x_j - \mu)$. Proposition 9.4 shows that $p_1, \ldots, p_n$ are independent. A change of variables gives

$$\tfrac{1}{\sqrt{2\pi}}\, e^{-p_j^2/2}\, dp_j = \tfrac{1}{\sigma\sqrt{2\pi}}\, e^{-(x_j-\mu)^2/(2\sigma^2)}\, dx_j,$$

and it follows that p_j has probability distribution $N(0, 1)$. Next define random variables $y_1, \ldots, y_n$ by $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = A \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$. Lemma 9.24 shows that $y_1, \ldots, y_n$ are independent and that each has $N(0, 1)$ as probability distribution. The first of the new random variables is

$$y_1 = \tfrac{1}{\sqrt{n}}(p_1 + \cdots + p_n) = \tfrac{1}{\sigma\sqrt{n}}\,(x_1 + \cdots + x_n - n\mu) = \tfrac{\sqrt{n}}{\sigma}(\bar{x} - \mu) = w. \quad (*)$$

In particular, w has probability distribution $N(0, 1)$. We calculate that

$$s^2 = \tfrac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \tfrac{1}{n-1}\Big(\sum_{j=1}^{n} x_j^2 - 2\bar{x}\sum_{j=1}^{n} x_j + n\bar{x}^2\Big) = \tfrac{1}{n-1}\Big(\sum_{j=1}^{n} x_j^2 - n\bar{x}^2\Big). \quad (**)$$

Consequently

$$\begin{aligned}
\sum_{j=2}^{n} y_j^2 &= \sum_{j=1}^{n} y_j^2 - \sigma^{-2}n(\bar{x} - \mu)^2 && \text{by } (*)\\
&= \sum_{j=1}^{n} \big(\tfrac{x_j-\mu}{\sigma}\big)^2 - \sigma^{-2}n(\bar{x} - \mu)^2 && \text{since } A \text{ is orthogonal}\\
&= \Big(\sigma^{-2}\sum_{j=1}^{n} x_j^2 - 2\sigma^{-2}\mu n\bar{x} + \sigma^{-2}n\mu^2\Big)\\
&\quad + \big(-\sigma^{-2}n\bar{x}^2 + 2\sigma^{-2}n\bar{x}\mu - \sigma^{-2}n\mu^2\big)\\
&= \sigma^{-2}\sum_{j=1}^{n} x_j^2 - \sigma^{-2}n\bar{x}^2\\
&= \sigma^{-2}(n-1)s^2 && \text{by } (**)\\
&= v.
\end{aligned}$$

[25]This particular choice of A is called a **Helmert matrix**. Other choices can be obtained by extending the first row of the above B to a basis of row vectors and then applying the Gram–Schmidt orthogonalization process to the rows as in *Basic*, digital second edition, page 599. All such choices will serve in the present proof.

Since the y_j are independent and each has probability distribution $N(0, 1)$, $v = \sigma^{-2}(n-1)s^2$ is exhibited as a sum of squares of independent random variables each distributed according to $N(0, 1)$, and it has probability distribution $\chi^2(n-1)$. Since w depends only on y_1 and v depends only on $y_2, \dots, y_{n-1}$, Proposition 9.4 shows that w and v are independent. Lemma 9.23 is therefore applicable with $n = k - 1$, and Theorem 9.20 follows. □

BIBLIOGRAPHICAL REMARKS ABOUT CHAPTER IX. The proof of Theorem 9.5 is adapted from Doob's *Measure Theory*, and the proof of Theorem 9.8 is adapted from Feller's Volume II of *An Introduction to Probability Theory and Its Applications*. The material in Sections 5–9 is partly adapted from the Wikipedia article "Central Limit Theorem," as of January 2015, and from the chapter[26] "Limit Theorems" of the online *Analysis of Data* project of Guy Lebanon dated 2012. The mathematics in Section 10 is adapted from the chapter[27] "Distributions Derived from the Normal Distribution" in the online lecture notes of Barbara Bailey on multivariate statistics from Summer 2009. The organization and overview of Section 10 and the end of Section 9 owe much to conversations with two statisticians, Sarah Knapp Abramowitz and Jon Kettenring.

11. Problems

1. If x is a random variable with probability distribution μ_x, find a formula for the probability distribution $\mu_{|x|}$ of $|x|$ in terms of μ_x.
2. Let $x_1, \dots, x_N$ be random variables on a probability space (Ω, P), let $\mu_{x_1,\dots,x_N}$ be their joint probability distribution, and let $\Phi : \mathbb{R}^N \to \mathbb{R}$ be a nonnegative Borel function. Prove that
$$\int_{\mathbb{R}} \Phi(t_1, \dots, t_N)\, d\mu_{x_1,\dots,x_N}(t_1, \dots, t_N) = \int_{\mathbb{R}} s\, d\mu_{\Phi\circ(x_1,\dots,x_N)}(s),$$
where $\mu_{\Phi\circ(x_1,\dots,x_N)}$ is the probability distribution of $\Phi \circ (x_1, \dots, x_N)$.
3. Suppose on a probability space (Ω, P) that $\{y_n\}_{n=1}^{\infty}$ is a sequence of random variables with a common mean μ and with variance σ_n^2, and suppose that $\Phi : \mathbb{R} \to \mathbb{R}$ is a bounded continuous function.
   (a) Prove that $P(\{|y_n - \mu| \geq \delta\}) \leq \sigma_n^2 \delta^{-2}$ for all n.
   (b) Suppose that $|\Phi| \leq M$ and that δ and ϵ are positive numbers such that $|t - \mu| < \delta$ implies $|\Phi(t) - \Phi(\mu)| < \epsilon$. Prove that $|E(\Phi(y_n)) - \Phi(\mu)| \leq \epsilon + 2M\sigma_n^2\delta^{-2}$.
   (c) Prove that if $\lim_n \sigma_n^2 = 0$, then $\lim_n E(\Phi(y_n)) = \Phi(\mu)$.

[26] This is Chapter 8 of Volume 1.

[27] This is Chapter 6, and the relevant lectures are Lectures 9 and 11.

(d) Show that the argument in (c) continues to work if Φ is the indicator function of an interval whose closure does not contain μ. Why does the conclusion in this case contain the conclusion of the Weak Law of Large Numbers as in Theorem 9.7?

4. **(Bernstein polynomials)** This problem gives a constructive proof of the Weierstrass Approximation Theorem by using probability theory.
   (a) Fix p with $0 \le p \le 1$. A certain unbalanced coin comes up "heads" with probability p and "tails" with probability $1 - p$; "heads" is scored as the outcome 1, and "tails" is scored as the outcome 0. Set up a probability model (Ω, P) for a sequence of independent coin tosses of this unbalanced coin, and let x_n be the outcome of the n^{th} toss.
   (b) Show that the mean of the outcome of a single toss of the coin is p and the variance is $p(1 - p)$.
   (c) Let $s_n = x_1 + \cdots + x_n$. Show for each integer k with $k \le n$ that $P(\{s_n = k\}) = \binom{n}{k} p^k (1 - p)^{n-k}$.
   (d) For continuous $\Phi : [0, 1] \to \mathbb{R}$, extend Φ to all of $\mathbb{R}$ so as to be constant on $(-\infty, 0]$ and on $[1, +\infty)$. Apply the result of Problem 3c to show that $\lim_n \sum_{k=0}^n \Phi\left(\frac{k}{n}\right)\binom{n}{k} p^k (1 - p)^{n-k} = \Phi(p)$.
   (e) Prove that the convergence in (d) is uniform for $0 \le p \le 1$, and conclude that Φ is the uniform limit of an explicit sequence of polynomials on $[0, 1]$.

Problems 5–9 are closely related to the Kolmogorov Extension Theorem (Theorem 9.5) and in a sense explain the mystery behind its proof. Let X be a compact metric space, and for each integer $n \ge 1$, let X_n be a copy of X. Define $\Omega^{(N)} = \bigtimes_{n=1}^{N} X_n$, and let $\Omega = \bigtimes_{n=1}^{\infty} X_n$. Each of $\Omega^{(N)}$ and Ω is given the product topology. If E is a Borel subset of $\Omega^{(N)}$, we can regard E as a subset of Ω by identifying E with $E \times \left(\bigtimes_{n=N+1}^{\infty} X_n\right)$. In this way any Borel measure on $\Omega^{(N)}$ can be regarded as a measure on a certain σ-subalgebra $\mathcal{F}_N$ of the σ-algebra $\mathcal{B}(\Omega)$ of Borel sets.

5. Prove that $\bigcup_{n=1}^{\infty} \mathcal{F}_n = \mathcal{F}$ is an algebra of sets.

6. Let ν_n be a (regular) Borel measure on $\Omega^{(n)}$ with $\nu(\Omega^{(n)}) = 1$, and regard ν_n as defined on $\mathcal{F}_n$. Suppose for each n that ν_n agrees with ν_{n+1} on $\mathcal{F}_n$. Define $\nu(E)$ for E in $\mathcal{F}$ to be the common value of $\nu_n(E)$ for n large. Prove that ν is nonnegative additive, and prove that in a suitable sense ν is regular on $\mathcal{F}$.

7. Using the kind of regularity established in the previous problem, prove that ν is completely additive on $\mathcal{F}$.

8. In view of Problems 6 and 7, ν extends to a measure on the smallest σ-algebra for Ω containing $\mathcal{F}$. Prove that this σ-algebra is $\mathcal{B}(\Omega)$.

9. Let X be a 2-point space, and let ν_n be 2^{-n} on each one-point subset of $\Omega^{(n)}$, so that the resulting ν on Ω is coin-tossing measure on the space of all sequences of "heads" and "tails." Exhibit a homeomorphism of Ω onto the standard Cantor set in $[0, 1]$ that sends ν to the usual Cantor measure, which is the Stieltjes measure corresponding to the Cantor function that is constructed in Section VI.8 of *Basic*.

Problems 10–14 concern the Kolmogorov Extension Theorem (Theorem 9.5) and its application to Brownian motion. If J is a subset of the index set I, a subset A of Ω will be said to be of type J if A can be described by

$$A = x_J^{-1}(E) = \{\omega \in \Omega \mid x_J \in E\} \qquad \text{for some subset } E \subseteq S^J.$$

As in the statement of the Kolmogorov theorem, let $\mathcal{A}'$ be the smallest algebra containing all subsets of Ω that are measurable of type F for some finite subset F of I. Let $\mathcal{A}$ be the smallest σ-algebra containing $\mathcal{A}'$.

10. From the fact that the collection of subsets of Ω that are of type J is a σ-algebra, prove that every set in $\mathcal{A}$ is of type J for some countable set J.
11. Form Brownian motion for time $I = [0, T]$ by means of the Kolmogorov Extension Theorem. Let C be the subset of continuous elements ω in Ω. Prove that C is not in $\mathcal{A}$.
12. With C as in Problem 11, prove that the only member of $\mathcal{A}$ contained in C is the empty set, and conclude that the inner measure of C relative to P is 0.
13. Still with C as in Problem 11, suppose that E is a subset of Ω of type J for some countable J and that $C \subseteq E$. Prove that the set C_J of elements ω in Ω that are uniformly continuous on J is contained in E.
14. Still with C as in Problem 11, suppose for every countable subset J of I that the set C_J of elements ω in Ω that are uniformly continuous on J is in $\mathcal{A}$ and has $P(C_J) = 1$. Prove that the outer measure of C relative to P is 1.

Problems 15–19 concern methods of convergence and examples of them. In each problem, all random variables are assumed to be defined on a fixed probability space (Ω, P).

15. Prove that if a sequence of random variables $\{x_n\}$ converges to x in probability, then a subsequence of $\{x_n\}$ converges to x almost surely.
16. Suppose that c is a constant and $\{x_n\}$ is a sequence of random variables such that $E(x_n) = c$ for all n and $\lim_n \operatorname{Var}(x_n) = 0$. Prove that $\{x_n\}$ converges to the constant c in probability.
17. In connection with the implication (b) implies (c) in the Portmanteau Lemma, give an example to show that there exist a sequence $\{\mu_n\}$ of finite Stieltjes measures and another Stieltjes measure μ such that $\lim_n \int_{\mathbb{R}} g\, d\mu_n = \int_{\mathbb{R}} g\, d\mu$ for all $g \in C_{\text{com}}(\mathbb{R})$ but not for all bounded continuous g on $\mathbb{R}$.

18. **(Theorem of de Moivre and Laplace)** The setting of this problem is the same as for Problems 4a–c, the repeated tossing of an unbalanced coin, and the result is the original historical conclusion of the Central Limit Theorem. Using Theorem 9.19, verify that
$$\lim_n P\Big(\omega \,\Big|\, a < (\sqrt{n})^{-1}\big(s_n(\omega) - np\big) < b\Big) = \frac{\int_a^b \exp\left(-\frac{u^2}{2p(1-p)}\right) du}{\sqrt{2\pi p(1-p)}}.$$

19. For integers $k \geq 0$ and $n \geq 0$ and real λ, define $p_{n,\lambda}(k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k}$ and $p_\lambda(k) = \frac{\lambda^k}{k!}\, e^{-\lambda}$.
    (a) Check that $\sum_{k=0}^{n} p_{n,\lambda}(k) = 1$ and that $\sum_{k=0}^{\infty} p_\lambda(k) = 1$. The probability distribution that assigns weight $p_{n,\lambda}(k)$ to the integer k is called the **binomial distribution** with parameters n and λ, and the probability distribution that assigns weight $p_\lambda(k)$ to the integer k is called the **Poisson distribution** with parameter λ.
    (b) Fix λ, let $x_{n,\lambda}$ be a random variable having probability distribution given by $p_{n,\lambda}$, and let x_λ be a random variable having probability distribution given by p_λ. Prove that $\{x_{n,\lambda}\}$ converges to x_λ in distribution.
    (c) Calculate the mean and variance of x_λ.

20. In Theorem 9.20 write $f_n(t) = c_n\left(1+\frac{t^2}{n-1}\right)^{-n/2}$, i.e., write c_n for the coefficient in the statement of the theorem. This problem shows that $f_n(t)$ converges pointwise to $(2\pi)^{-1/2}e^{-t^2/2}$, as was asserted just before Figure 9.3.
    (a) Prove for arbitrary $c > 0$ that $s \mapsto (1 + c/s)^s$ is an increasing function of s for $s > 0$.
    (b) Deduce from (a) that $n \mapsto \left(1+\frac{t^2}{n-1}\right)^{-(n-1)/2}$ decreases to $e^{-t^2/2}$ as n tends to $+\infty$, and explain why it follows that $\lim_{n\to\infty} c_n^{-1} f_n(t) = e^{-t^2/2}$ pointwise.
    (c) Using the Dominated Convergence Theorem, prove the pointwise limit formula
$$\lim_{n\to\infty} \int_{\mathbb{R}} \left(1 + \frac{t^2}{n-1}\right)^{-n/2} dt = \int_{\mathbb{R}} e^{-t^2/2}\, dt,$$
    and deduce as a consequence that $\lim_{n\to\infty} c_n^{-1} = \sqrt{2\pi}$.

21. In the setting in Section 10 of sample size n, let $t_n = (\bar{x} - \mu)/(s/\sqrt{n})$ be the random variable defined before the statement of Theorem 9.20, and let t_∞ be a random variable with distribution $N(0, 1)$. Suppose that the sample size n in Theorem 9.20 is allowed to tend to infinity. Explain how it follows from Problem 20 that the random variables t_n converge to t_∞ in distribution.

Problems 22–26 give a direct computational proof, without characteristic functions, of the Central Limit Theorem (Theorem 9.19) under the assumption that the common distribution of the random variables x_n is normal of type $N(\mu, \sigma^2)$.

22. If c is a constant and if a random variable x has a probability distribution of the form $f(t)\,dt$, what is the form of the probability distribution of $x + c$?

23. If c is a positive constant and if a random variable x has a probability distribution of the form $f(t)\,dt$, what is the form of the probability distribution of cx?

24. If x and y are independent random variables whose respective probability distributions are of the form $f(x)\,dx$ and $g(x)\,dx$, it was shown in Example 3b in Section 2 that the probability distribution of $x + y$ is $(f * g)(x)\,dx$. Under the assumption that x and y are independent and have probability distributions both normal, of the respective forms $N(\mu, \sigma^2)$ and $N(\mu', \sigma'^2)$, show that $x + y$ is normally distributed of the form $N(\mu + \mu', \sigma^2 + \sigma'^2)$.

25. Suppose that $\{x_n,\ n \geq 1\}$ is a sequence of independent random variables, each with the probability distribution $N(\mu, \sigma^2)$. Find the distribution of the random variable
$$w_n = \sqrt{n}\left(\frac{x_1 + \cdots + x_n}{n} - \mu\right).$$

26. In what sense is the convergence in distribution of the random variables w_n in the previous problem to a random variable with probability distribution $N(0, \sigma^2)$ trivial?