AppC.tex — January 14, 2000

P.J.M. Bongaarts
Instituut Lorentz, University of Leiden
bongaart@lorentz.leidenuniv.nl

Topics from 20^{th} century physics.
An introductory course for students in mathematics

APPENDIX C: PROBABILITY THEORY

1. Introduction
    1.1. History
    1.2. Probability theory in physics
2. Kolmogorov's formulation of probability theory
    2.1. Discrete probability theory
    2.2. The general framework
3. Mathematical properties of distribution functions
    3.1. Introductory remark
    3.2. Distribution functions in a single variable
    3.3. Distribution functions in n variables
    3.4. Stochastic variables as a bacic notion

1. INTRODUCTION

1.1. *History*

Probability theory is a set of ideas and procedures which allows us to act reasonably in situations where we have incomplete information. It arose, as a defnite mathematical discipline, in the 17^{th} century, its main applications being gambling and insurance. It had to wait untill the first half of the 20^{th} century before its general principles were put together in a rigorous mathematical formalism. This was provided in the thirties by the Russian mathematician Kolmogorov; his 'axiomatic' formulation in terms of the theory of measure and integration and, more generally, functional analysis, has become the generally accepted standard form of probability theory.

1.2. *Probability theory in physics*

Probability theory has played an important role in physics since the second half of the 19^{th} century, when it was realized that matter is composed of a very large number of separate particles, atoms and molecules, and that therefore the properties of systems of gases and liquids could to a large extent be understood from averaging in some statistical way over the behaviour of the individual particles. In this connection one should mention the names of Ludwig Boltzmann (1844-1906), James Clerk Maxwell (1831-1879), also known for his fundamental theory of electromagnetism, and above all Josiah Willard Gibbs (1839-1903), the first American theoretical physicist of international fame, who laid the foundations of what is now known as *statistical mechanics*. He was working before the mathematical formalization of probability theory by Kolmogorov; some of the terminology he introduced for basic probabilistic notions has survived in physic only. A good example is the notion of *ensemble*; a mathematician reading a modern physics text book will need some time to realize that this is just a probability measure.

A further use of probability theory in physics came with quantum mechanics. This is in fact the reason that this appendix is incuded in these notes. Quantum theory implies a fundamental lack of determinism in physics. Philosophers of science do not agree on what this precisely means, even though there are no problems at the level of applications to physical situations. In any case the use of probability theory in quantum mechanics is much more intriguing and fundamental than in the rest of physics. As a mathematical model quantum theory can be regarded as a non-trivial generalization of 'classical' probability theory.

2. KOLMOGOROW'S FORMULATION

2.1. *Discrete probability theory*

Before setting up the general formalism it is useful to consider briefly the simple case in which one assigns probabilities to a finite number of n elementary events, for example the 6 sides of a dice. Let such possible outcomes of an experiment, be denoted by the integers $1, \ldots, n$. The probability that the p^{th} event turns up in an experiment should be equal to a nonnegative real number ρ_p, with of course $\sum_{p=1}^{n} \rho_p = 1$. For a given function $f(p)$ of the outcomes, on defines the the average value of the 'random value' f as $\sum_{p=1}^{n} f(p)\rho_p$. The standard interpretation of such probabilities and averages is based on the idea of a repeating the same experiment many times, but that does not concern us here. One may ask for the probability that any one of a subset of the basic events $1, \ldots, n$ turns up. The answer is of course the sum of the numbers ρ_p with the p belonging to the subset. It is therefore sensible to call all the subsets of $1, \ldots, n$ *events*. This is a good stepping stone for going to 'continuous' situations, say that of throwing a dart arrow instead of a dice. The probability of hitting a given point inside a circle is not a useful concept, as it is clearly 0. To have a nonzero probability one needs as events sets with a nonzero volume. This leads to probabilities as *measures* of sets, in the technical sense of measure theory; averages become integrals with respects to this measure. This is precisely the starting point of Kolomogorov's 'axiomatic' formulation of probability theory .

2.2. *The general framework*

Consider a *measure space*, i.e. a triple $(\Omega, \mathcal{B}, \mu)$, consisting of:

a. A nonempty set Ω .

b. A σ-algebra $\mathcal{B}$, i.e. a system of subsets of Ω such that it contains the empty set and Ω itself, with every subset its complement and with every denumerable system of subsets its union and intersection.

c. A measure μ, i.e. a function on the sets in $\mathcal{B}$, which assigns the number 0 to the empty set and further to each subset in $\mathcal{B}$ a nonnegative real number or possibly $+\infty$, such that for A and B in $\mathcal{B}$ with $A \subset B$ one has $\mu(A) \leq \mu(B)$ and such that for a denumerable system $A_1, A_2, \ldots$ of pairwise disjoint subsets in $\mathcal{B}$ one has $\mu(A_1 \cup A_2 \cup \ldots) = \sum_{j=1} \mu(A_j)$.

If we assume $\mu(\Omega) = 1$, then $(\Omega, \mathcal{B}, \mu)$ is called a *probability measure space*. In this case we use the notation P (from probability) instead of μ.

A function f from Ω to the real numbers, possibly with $-\infty$ and $+\infty$ added, is called a *measurable function*, if for each real λ the inverse image $f^{-1}((-\infty, \lambda])$ belongs to $\mathcal{B}$.

DEFINITION: A *probability theory* is a probability measure space $(X, \mathcal{B}, \mu)$. The subsets in $\mathcal{B}$ are called *events*, with X the *certain event* and the empty set the *impossible event*. The real number $\mu(A)$, nonnegative and smaller or equal to 1, is the *probability* for the event A.

DEFINITION: A (real-valued) *random variable* or *stochastic variable* with respect to this probability theory is a real-valued measurable function on Ω. Its average or *expectation* is defined as $E(f) = \int_{\Omega} f(\omega)P(d\omega)$. This expectation exists if the integral is finite. More generally, $E(f^p)$, when it exists, is called the p^{th} *moment* of f. In the applications in the main text of these notes the expectation $E(f)$, – or 'expectation value' as it is called in physics – will often be denoted by $\overline{f}$ or $< f >$.

DEFINITION: The expression $E\left(((f - E(f))^2\right) = E(f^2) - (E(f))^2$ is called the *variance* of f; its square root is usually denoted as σ or $\sigma(f)$ – in physical applications as Δf – and called the *standard deviation*.

DEFINITION: A stochastic variable f has a *distribution function*. We denote it as F. For each real x the value $F(x)$ is defined as the probability that the random variable takes a value smaller or equal to x, i.e.

$$F(x) := P(f^{-1}((-\infty, x]) = P(\{\omega \in \Omega \,|\, f(\omega) \leq x\}).$$

One often consider *systems* $\{f_\alpha\}$ of random variables. Such systems can be arbitrary; a system of random variables indexed by a continuous parameter t, usually time, is called a *stochastic process*. Here we restrict the discussion to finite systems $f_1, \ldots, f_n$. Such a system has a *joint* or *simultaneous distribution function*. It is a real valued function on R^n defined simply as

$$F(x_1, \ldots, x_n) := P\left(\{\omega \in \Omega \,|\, f_1(\omega) \leq x_1, \ldots, f_n(\omega) \leq x_n\}\right).$$

DEFINITION: The expression $E\left((f - E(f))(g - E(g))\right)$ is denoted as $\mathrm{Cov}(f, g)$ and called the *covariance* of f and g. Its normalized form $\mathrm{Cov}(f, g)/\sigma(f)\sigma(g)$ is called the *coefficient of correlation*; it takes values between $-1$and $+1$. The random variables f and g are called *uncorrelated* if $\mathrm{Cov}(f, g) = 0$.

3. PROPERTIES OF DISTRIBUTION FUNCTIONS

3.1. *Introductory remark*

It is worthwhile to discuss the mathematical properties of distribution functions in somewhat more detail. We consider first the case of distribution functions of a single variable, and then the general case of n variables.

3.2. *Distribution functions of a single variable*

One verifies easily that a distribution function associated with a random variable f, as defined in the preceding chapter, is a real-valued, monotone nondecreasing function F on the real line, continuous from the right, with moreover $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$. Denote, just for this subsection, the collection of such functions as $\mathcal{F}$. One proves easily that $\mathcal{F}$ is in one-to-one correspondence with the collection of probability measures on the real line, with with as σ-algebra of subsets the system of *Borel* sets, the smallest σ-algebra containing all open sets in R. On one hand one defines, for a given function F, a probability measure P_F by extension of the formula $P_F((x_1, x_2]) = F(x_2) - F(x_1)$, for all half-open intervals $(x_1, x_2]$, on the other hand one obtains from a probability measure μ on R a function F_μ as $F_\mu(x) = \mu((-\infty, x])$. These formulas mean that each F is the distribution function for the random variable x with respect to a unique probability theory $(\Omega, \mathcal{B}, P_F)$ with $\Omega = R$, $\mathcal{B}$ the Borel σ-algebra of R and P_F an unique probability measure. Because of this it is reasonable to use the term *distribution function* for arbitrary functions in $\mathcal{F}$. Note that for a distribution function in this sense monotonicity implies the existence, for each x in R, of $F_-(x) := \sup_{x'<x} F(x')$ and $F_+(x) := \inf_{x'>x} F(x')$, this without using the right continuity. One has of course $F_-(x) \leq F_+(x)$, for all x in R. It should also be remarked that it is clear from this that continuity from the *right* in this set up is a matter of convention.

Consider the union of all open intervals (x_1, x_2) on which a given distribution function is constant. Its complement is a nonempty closed set in R, the *support* of F, denoted as $\operatorname{supp} F$. It is also the support of the probability measure in R associated with F. Note that the support of F can equivalently be defined as $\{x \in R \,|\, F(x-\varepsilon) < F(x+\varepsilon), \forall \varepsilon > 0\}$. Points x for which $F_-(x) < F_+(x)$ obviously belong to the support of F. They are called the *discrete points* of $\operatorname{supp} F$. If the support of F has only disrete points, i.e. if the function F is a stepfunction, we are back in discrete probability theory. For points x with $F_-(x) = F_+(x)$ the function F is continuous in x. If F is not only continuous but *absolutely continuous* , a stronger requirement, then F is differentiable, except in a set of point with Lebèsgue measure 0; the derivative is an integrable function ρ, nonnegative, with integral $\int_{-\infty}^{+\infty} \rho(x)dx = 1$; its is called a *probability density*. In physics one usually considers probabilities to be discrete or to be given by probability densities. Not much is lost by this simplification.

3.3. *Distribution functions in n variables*

For a system $f_1, \ldots, f_n$ of random variables the distribution function is a real-valued function on R^n, monotone nondecreasing and continuous from the right in each variable x_j, with moreover $\lim_{x_j \to -\infty} F(x_1, \ldots, x_j, \ldots, x_n) = 0$ and $\lim_{x_j \to +\infty, j=1,\ldots,n} F(x_1, \ldots, x_n) = 1$ for each x_j. The distribution function F_{j_0}, the 1-variable distribution function of the single random variable f_{j_0}, can be obtained from the n-variable function as a *marginal* distribution function by taking the limit $x_j \to 1$, for all $j \neq j_0$. In this manner all the information on the f_j as separate random variables is contained in the n-variable function; the converse is of course not true. Functions $F(x_1, \ldots, x_n)$ with all these properties can be called n-variable distribution functions, for a reason similar to the 1-variable case: there is a one-to-one correspondence between such functions and probability measures on R^n. Each F can then be regarded as the n-variable distribution function associated with the n random variables $x_1, \ldots, x_n$ with respect to a probability theory $(\Omega, \mathcal{B}, P_f)$, with $\Omega = R^n$, $\mathcal{B}$ the Borel σ-algebra of R^n and P_F a unique probability measure. One can again define the notion of support of a distribution function and correspondingly of the associated probability measure on R^n, with similar formulas. This is left to the reader.

3.4. *Stochastic variables as a bacic notion*

The general definition of a stochastic (or random) variable is a measurable function on a probabilty measure space. In working with stochastic variables this measure space is often not mentioned let alone explicitly described; one just assumes that it exists and is present somewhere in the background. What is given is the joint distribution function of the variables. As long as one has a fixed set of stochastic variables in mind this is all one needs; the distribution function contains all information. It is not hard to show, for instance for a single stochastic variable f, using the results from the preceding subsections, that the expectation of f, defined in terms of the underlying probability space $(\Omega, \mathcal{B}, P)$ as $E(f) = \int_\Omega f(\omega)P(d\omega)$, can be written as a Lebèsgue-Stieltjes integral over F as $\int_{-\infty}^{+\infty} x dF(x)$. In the case of absolute continuity, with a probability density ρ, this becomes the ordinary Lebèsgue integral $\int_{-\infty}^{+\infty} x\rho(x)dx$. One has similarly for a system $f_1, \ldots, f_n$ of stochastic variables with n-variable distribution function $E(f_j) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_j dF(x_1, \ldots, x_n)$, for $j = 1, \ldots, n$, which is of course equal to the 1-dimensional integral $\int_{-\infty}^{+\infty} x_j dF_j(x_j)$, with F_j the marginal distribution function for f_j. The covariance of two variables f_j and f_k, another important quantity, defined as an integral over Ω with respect to P, becomes a Lebèsgue-Stieltjes integral over R^n, which can be reduced immediately to a two dimensional integral. All this means that for a fixed system of n stochastic variables one can, instead of the original probability space $(\Omega, \mathcal{B}, P)$ which is supposed to be in the background, effectively use an explicit probability theory in R^n, determined by a joint distribution function $F(x_1, \ldots, x_n)$ in the manner described in the preceding sections. It finally means that for practical purposes a system of n random variables can be *defined* as a n-variable distribution function F on R^n.