Notes on Probability Theory
Module VII of Refresher Course on Theoretical Physics
17-31 March, 2025, Sri S. Ramasamy Naidu Memorial College, Sattur, Tamil Nadu

Govind S. Krishnaswami, Chennai Mathematical Institute March 23, 2025

Please let me know at govind@cmi.ac.in of any comments or corrections

Contents

1 References

1. Y A Rozanov, *Probability Theory: A Concise Course*, Rev. English Edition, Translated and Edited by R. A. Silverman, Dover (1977).
2. W Feller, *An Introduction to Probability Theory and its Applications*, John Wiley, Vol I, 3rd Ed. (1968), Vol II, (1966).
3. V Balakrishnan, Mathematical Physics with applications, problems and solutions, ANE Books (2020).

2 Events, sample space, probabilities and combination of events

Outcomes of an experiment. Suppose the performance of an experiment can lead to various mutually exclusive outcomes. E.g.: (a) Tossing a coin can lead to the outcomes heads and tails. (b) Rolling a die can lead to the outcomes 1,2,3,4,5,6 displayed on the die. (c) Switching on a digital thermometer can lead to outcomes consisting of temperature estimates such as 24.1, 22.5. 23.0 Celsius etc.

Elementary events are the possible mutually exclusive outcomes of an experiment. The typical elementary event is denoted ω. The set of possible elementary events of a given experiment is called the **sample space**, denoted Ω.

General events. Event A is associated to the elementary events of an experiment if given any elementary event, we can say whether or not the outcome ω leads to the occurrence of A. For example, event A could be that the outcome of rolling a die is even. In this case, there are three elementary events $\omega = 2, 4, 6$ that lead to the occurrence of A. Thus, a general event A is associated to a set of elementary events. Bearing this in mind, we may view an event A as simply a subset of the sample space Ω.

Probability of an event when outcomes are equally likely. Consider an experiment with a finite number N of mutually exclusive outcomes. This means the sample space Ω is a finite set of N elementary events. Suppose further that the elementary events are all equally likely. Then the probability of the event A is defined as the fraction of outcomes in which A occurs: $\mathbf{P}(A) = N(A)/N$ where $N(A)$ is the number of elementary outcomes leading to the occurrence of A. For example, in the rolling of a die, the event A corresponding to an even outcome has probability $\mathbf{P}(A) = 3/6$.

Frequentist definition of probability. Experience allows us to extend the notion of probability beyond experiments with equally likely outcomes. Suppose an experiment can be repeatedly performed resulting in a sequence of independent trials under the same conditions. These trials may be the repeated tossing of a coin or the repeated estimation of the direction of wind in the horizontal plane at a fixed point in a room. In each of these trials, we suppose that (based on chance) an event A of interest either occurs or does not occur. The event A could be, for instance, the occurrence of heads or the wind direction lying between North and North-East. Now, suppose the experiment is repeated n times and the event A occurs in $n(A)$ of the trials. Then the relative frequency of the event A in the given sequence of trials is $n(A)/n$. Remarkably, it is found that the relative frequencies arising in different sequences of n trials approach a common value as the number of trials grows indefinitely. This limiting frequency[1]

$$\mathbf{P}(A) = \lim_{n\to\infty} n(A)/n \tag{1}$$

is called the probability of the event A in the given experiment.

- Evidently, the probability of any event A must be a real number with $0 \le \mathbf{P}(A) \le 1$.
- A pair of events A_1 and A_2 are mutually exclusive or incompatible if they cannot both occur simultaneously. For example the events corresponding to an even and an odd outcome from the roll of a die are incompatible. Viewed as subsets of the sample space Ω, mutually exclusive events have empty intersection. An event A and its complement $\bar{A}$ (the event that A does not occur) are mutually exclusive.

Combining events. Unions and intersections of sequences of events are defined in a natural way. One often abbreviates $A \cap B = AB$. The difference between two events $A_1 \setminus A_2$ is one where A_1 occurs but A_2 does not. In particular $\bar{A} = \Omega \setminus A$.

Visualizing relations between events. It is convenient to represent the sample space Ω by a plane region whose points are the elementary events. Then events, which are subsets of Ω are represented by various subsets of the plane region. In particular, $\bar{A}_1$ is the complement of A_1. The complement of Ω is the empty set $\emptyset$ corresponding to the event that nothing happened. Mutually exclusive events are represented by disjoint subsets. Draw figures to illustrate the following general relations (i) If $A_1 \subset A_2$ then $\bar{A}_1 \supset \bar{A}_2$. (ii) If $A = A_1 \cup A_2$ then $\bar{A} = \bar{A}_1 \cap \bar{A}_2$ and (c) If $A = A_1 \cap A_2$ then $\bar{A} = \bar{A}_1 \cup \bar{A}_2$. More generally, given a relation between events, we may obtain an equivalent relation by replacing events by their complements and changing $\cup, \cap, \subset, \supset$ to $\cap, \cup, \supset, \subset$.

[1] This formula is further justified by the strong law of large numbers.

Addition law for probabilities. Suppose A_1 and A_2 are a pair of mutually exclusive events associated with the outcomes of a random experiment and let $A = A_1 \cup A_2$. Suppose the experiment is repeated n times resulting in a series of independent trials under identical conditions. Let $n(A_1)$, $n(A_2)$ and $n(A)$ be the number of trials in which A_1, A_2 and A occur. Since they are mutually exclusive, $n(A) = n(A_1) + n(A_2)$ whence

$$\frac{n(A)}{n} = \frac{n(A_1)}{n} + \frac{n(A_2)}{n}. \tag{2}$$

For large n, these relative frequencies approach limits which coincide with the corresponding probabilities. Thus, for mutually exclusive events,

$$\mathbf{P}(A_1 \cup A_2) = \mathbf{P}(A_1) + \mathbf{P}(A_2). \tag{3}$$

Similarly, if A_1, A_2, A_3 are mutually exclusive, then $A_1 \cup A_2$ is mutually exclusive of A_3. Applying the previous addition law twice,

$$\mathbf{P}(A_1 \cup A_2 \cup A_3) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3). \tag{4}$$

This addition law extends to n mutually exclusive events for $n = 2, 3, 4, \ldots$.

• $\mathbf{P}(\Omega) = 1$ and $\mathbf{P}(\emptyset) = 0$

3 Conditional probability and statistical independence

Conditional probability. This concerns how the occurrence of one event is influenced by that of another event. The probability of A occurring given that B is known to have occurred is denoted $\mathbf{P}(A|B)$ and may be expressed as

$$\mathbf{P}(A|B) = \mathbf{P}(AB)/\mathbf{P}(B), \tag{5}$$

assuming $\mathbf{P}(B) \neq 0$ (so we cannot take $B = \emptyset$). This may be understood by writing it as $\mathbf{P}(AB) = \mathbf{P}(B)\mathbf{P}(A|B)$. In other words, the probability that both A and B occur may be factorized as the product of the probability of B times the probability that A occurs given that B did. Evidently, one also has $\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B|A)$.

• Some consequences of the definition of conditional probabilities follow.

1. Since $\mathbf{P}(AB) \leq \mathbf{P}(B)$, conditional probabilities must lie in the interval $[0, 1]$
2. If A and B are mutually exclusive, $\mathbf{P}(A|B) = 0 = \mathbf{P}(B|A)$.
3. If B implies A so that $B \subset A$, then $\mathbf{P}(A|B) = 1$.
4. Suppose $A_1, A_2, \cdots$ are mutually exclusive events with (disjoint) union $A = \cup_k A_k$. Then
$$\mathbf{P}(A|B) = \sum_k \mathbf{P}(A_k|B). \tag{6}$$
5. Suppose $A_1, A_2, \cdots$ is an exhaustive collection of mutually exclusive events in the sense that precisely one of the A_k always occurs ($\cup_k A_k = \Omega$), then
$$\mathbf{P}(A) = \sum_k \mathbf{P}(A|A_k) \quad \text{for any event } A. \tag{7}$$

   This formula is often helpful in calculating $\mathbf{P}(A)$.

Statistical independence. Two events A_1 and A_2 are said to be statistically independent or simply independent if the probability that both occur factorizes as a product:

$$\mathbf{P}(A_1A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2). \tag{8}$$

If this factorization does not hold, the events are statistically dependent.

This definition is motivated by the intuitive idea that A_1 and A_2 are independent, if the occurrence of A_2 has no bearing on the probability of occurrence of A_1 and vice versa. In terms of conditional probabilities, this is the assertion that

$$\mathbf{P}(A_1|A_2) = \mathbf{P}(A_1) \quad \text{and} \quad \mathbf{P}(A_2|A_1) = \mathbf{P}(A_2) \tag{9}$$

It follows that $\mathbf{P}(A_1A_2)/\mathbf{P}(A_2) = \mathbf{P}(A_1)$.

Statistical independence generalizes to several events. $A_1, \cdots, A_n$ are mutually independent if the probability that any q (for $2 \le q \le n$) of them to occur simultaneously factorizes as a product of individual probabilities:

$$\begin{aligned} \mathbf{P}(A_iA_j) &= \mathbf{P}(A_i)\mathbf{P}(A_j), \quad \mathbf{P}(A_iA_jA_k) = \mathbf{P}(A_i)\mathbf{P}(A_j)\mathbf{P}(A_k), \cdots, \\ \mathbf{P}(A_1A_2\cdots A_n) &= \mathbf{P}(A_1)\mathbf{P}(A_2)\cdots\mathbf{P}(A_n), \end{aligned} \tag{10}$$

for all combinations of indices such that $1 \le i < j < \cdots \le n$. Notice that mutual independence is a stronger condition than pairwise statistical independence.

4 Random variables and probability distributions

Random variable. Suppose Ω is a sample space of events. A real random variable ξ is a function that assigns a real number to each elementary event $\omega \in \Omega$. For a coin toss, the function $\xi(\text{heads}) = 1, \xi(\text{tails}) = -1$ is an example of a random variable. Of course, there are other random variables such as $\eta(\text{heads}) = 23.45$ and $\eta(\text{tails}) = -\pi/7$.

Probability distribution of a random variable. Let $\mathbf{P}\{x' \le \xi \le x''\}$ be the probability of the event that ξ lies in the interval $[x', x'']$. Knowledge of $\mathbf{P}\{x' \le \xi \le x''\}$ for all $x' \le x''$ is said to characterize the probability distribution of the random variable ξ.

Discrete random variable. A random variable ξ is discrete (or has a discrete distribution) if it takes only a finite or denumerably infinite number of distinct values x with probabilities

$$P_\xi(x) = \mathbf{P}\{\xi = x\}, \tag{11}$$

subject to the condition that

$$\sum_{i \in I} P_\xi(x_i) = 1. \tag{12}$$

Here x_i (for i in some index set I) denote all the possible values taken by ξ. For the above random variable associated to a fair coin, $P_\xi(\pm 1) = \frac{1}{2}$.

• For a discrete random variable,

$$\mathbf{P}\{x' \le \xi \le x''\} = \sum_{i \in I,\, x' \le x_i \le x''} P_\xi(x_i). \tag{13}$$

• The probability density function (PDF) for a discrete random variable ξ that takes the values x_i with probabilities $P_\xi(x_i)$, is defined as a sum of Dirac delta functions

$$p_\xi(x) = \sum_{i \in I} P_\xi(x_i)\delta(x - x_i). \tag{14}$$

The utility of this definition is that the above sums of probabilities may be written as integrals:

$$\mathbf{P}\{x' \leq \xi \leq x''\} = \int_{-\infty}^{\infty} p_\xi(x)\, dx. \tag{15}$$

Such formulae then apply also to continuous random variables which we turn to next.

Continuous random variable and probability density. The random variable ξ is continuous (or has a continuous distribution) if

$$\mathbf{P}\{x' \leq \xi \leq x''\} = \int_{x'}^{x''} p_\xi(x)\, dx, \tag{16}$$

where the probability density function $p_\xi(x)$ is a nonnegative function with unit integral

$$\int_{-\infty}^{\infty} p_\xi(x)\, dx = 1. \tag{17}$$

Assuming that $p_\xi(x)$ is a continuous function, the probability of the event $\xi = x$ is zero: $\mathbf{P}\{\xi = x\} = 0$ since it is the integral of a continuous (and hence bounded) function over an interval of zero length. However, the probability that ξ lies in a dx neighborhood of x is given by $\mathbf{P}\{\xi \in [x, x + dx]\} \sim p_\xi(x)\, dx$.

Cumulative distribution function. The (cumulative) distribution function (CDF) $\Phi_\xi(x)$ is defined as the probability of the event that $\xi \leq x$:

$$\Phi_\xi(x) = \mathbf{P}\{\xi \leq x\} \quad \text{for} \quad -\infty < x < \infty. \tag{18}$$

It follows from the definition that $\lim_{x\to\infty} \Phi_\xi(x) = 1$.

• For a discrete random variable, the cumulative distribution is a staircase function

$$\Phi_\xi(x) = \sum_{i\in I, x_i \leq x} P_\xi(x_i) = \sum_{i\in I} P_\xi(x_i)\theta(x - x_i) \tag{19}$$

where $\theta(x)$ is the unit step function, equal to 0 for $x < 0$ and 1 for $x \geq 0$. At the discrete values x_i taken by the random variable, $\Phi_\xi(x)$ jumps up by $P_\xi(x_i)$.

• For a continuous random variable,

$$\Phi_\xi(x) = \int_{-\infty}^{x} p_\xi(x')\, dx'. \tag{20}$$

Evidently, the cumulative distribution function is a nondecreasing function.

• The derivative of the cumulative distribution function is the probability density

$$\frac{d\Phi_\xi(x)}{dx} = p_\xi(x). \tag{21}$$

This formula also applies to a discrete random variable if we define the derivative of the unit step function $\theta(x)$ to be the Dirac delta function $\delta(x)$.

Mixture of discrete and continuous distributions. There are situations where a random variable is neither discrete nor continuous but a mixture of both. The density of states of a quantum system with partly discrete and partly continuous energy spectrum is of this sort. The corresponding probability density function is a sum of a continuous function ρ_ξ and a weighted sum of Dirac deltas:

$$p_\xi(x) = \rho_\xi(x) + \sum_{i \in I} P_\xi(x_i)\delta(x - x_i). \tag{22}$$

To qualify as a probability density, we must of course have

$$\int_{-\infty}^{\infty} \rho_\xi(x)\, dx + \sum_i P_\xi(x_i) = 1. \tag{23}$$

5 Joint distribution and independent random variables

Joint probability distribution and density. The joint probability distribution of a pair of discrete random variables ξ_1, ξ_2 is characterized by the probabilities

$$P_{\xi_1,\xi_2}(x_1, x_2) = \mathbf{P}\{\xi_1 = x_1, \xi_2 = x_2\}. \tag{24}$$

We may say that $\xi = (\xi_1, \xi_2)$ is a 2 dimensional vector-valued random variable. The probability of the event $(\xi_1, \xi_2) \in B$ where B is a subset of $\mathbb{R}^2$ is

$$\mathbf{P}\{(\xi_1, \xi_2) \in B\} = \sum_{(x_1,x_2)\in B} P_{\xi_1,\xi_2}(x_1, x_2). \tag{25}$$

The joint probabilities may be expressed in terms of conditional probabilities:

$$P_{\xi_1,\xi_2}(x_1, x_2) = P_{\xi_1|\xi_2}(x_1|x_2)P_{\xi_2}(x_2). \tag{26}$$

Here $P_{\xi_1|\xi_2}(x_1|x_2)$ is the probability that ξ_1 takes the value x_1 given that ξ_2 takes the value x_2.
• For a pair of continuous random variables ξ_1, ξ_2, by the joint probability density, we mean a function $p_{\xi_1,\xi_2}(x_1, x_2)$ such that the probability of any event of the form $(\xi_1, \xi_2) \in B$ is given by

$$\mathbf{P}\{(\xi_1, \xi_2) \in B\} = \iint_B p_{\xi_1,\xi_2}(x_1, x_2)\, dx_1 dx_2. \tag{27}$$

• The joint density of ξ_1 and ξ_2 can be expressed in terms of the conditional probability density:

$$p_{\xi,\eta}(x, y) = p_{\xi|\eta}(x|y)p_\eta(y). \tag{28}$$

Independent random variables. A family of random variables $\xi_1, \xi_2, \cdots, \xi_n$ is statistically independent if the events $x'_k \le \xi_k \le x''_k$ for $k = 1, 2, \cdots, n$ are independent for any $x'_k \le x''_k$. The infinite sequence of random variables $\xi_1, \xi_2, \cdots$ are statistically independent if $\xi_1, \xi_2, \cdots, \xi_n$ are independent for each $n = 2, 3, \cdots$.
• Recall that independent events were defined via factorization of probabilities. It follows that the joint probability distribution of a pair of independent random variables is such that

$$P_{\xi_1,\xi_2}(x_1, x_2) = P_{\xi_1}(x_1)P_{\xi_2}(x_2) \quad \text{and} \quad p_{\xi_1,\xi_2}(x_1, x_2) = p_{\xi_1}(x_1)p_{\xi_2}(x_2) \tag{29}$$

for discrete and continuous random variables ξ_1 and ξ_2 respectively.

Marginal distribution Suppose ξ and η are a pair of discrete random variables with joint probability distribution encoded in the probabilities $P_{\xi,\eta}(x,y)$. We wish to find the (marginal) probability distribution of one of them, say, ξ without reference to the value of η. Since we must account for all possible values of η, the marginal distribution of ξ is obtained by summing over all possible values of η

$$P_\xi(x) = \sum_{y_j} P_{\xi,\eta}(x, y_j). \tag{30}$$

If $P_{\xi,\eta}$ is normalized to have integral 1, then check that P_ξ is automatically normalized to one.
• Suppose the values of the probabilities $P_{\xi,\eta}(x,y)$ are written in a rectangular array with rows labelled by ξ and columns by η. Then the marginal probabilities of ξ and η are obtained by adding up the entries in each row or column. These sums are conventionally written along the margins of the paper on which the array is written down. This explains the name marginal distribution.
• Analogously, for a pair of continuous random variables ξ, η the marginal density of ξ is given by averaging over all possible values of η:

$$p_\xi(x) = \int p_{\xi,\eta}(x,y)\, dy. \tag{31}$$

• The same averaging procedure is used in the construction of the reduced density matrix of a subsystem by tracing over the remaining degrees of freedom in the system. The idea of a marginal distribution also finds use in the passage from the micro-canonical to canonical distributions in classical statistical mechanics.

Convolution: distribution of the sum of two independent random variables. Suppose ξ_1 and ξ_2 are a pair of independent continuous random variables with probability densities $p_{\xi_1}(x_1)$ and $p_{\xi_2}(x_2)$. Then the probability density of their sum $\eta = \xi_1 + \xi_2$ is given by the convolution

$$p_\eta(y) = \int_{-\infty}^{\infty} p_{\xi_1}(y-x) p_{\xi_2}(x)\, dx. \tag{32}$$

To see why this is the case, we begin by noting that on account of their independence, the joint probability density is given by $p_{\xi_1,\xi_2}(x_1,x_2) = p_{\xi_1}(x_1)p_{\xi_2}(x_2)$. It follows that the probability that η lies in the interval $[y', y'']$ is given by

$$\begin{aligned} \mathbf{P}\{y' \le \eta \le y''\} &= \iint_{y' \le x_1 + x_2 \le y''} p_{\xi_1}(x_1) p_{\xi_2}(x_2)\, dx_1\, dx_2 \\ &= \int_{y'}^{y''} dy \int_{-\infty}^{\infty} p_{\xi_1}(y-x) p_{\xi_2}(x)\, dx \end{aligned} \tag{33}$$

where we let $y = x_1 + x_2$ and denoted x_2 by x.

Uniform distribution. Suppose a point ξ is 'tossed at random' into the interval $[a,b]$. This means the probability of ξ falling in the subinterval $[x', x''] \subset [a,b]$ is independent of the location of the interval. In other words, this probability must be translation invariant: $\mathbf{P}\{\xi \in [x', x'']\} = \mathbf{P}\{\xi \in [x'+c, x''+c]\}$ for all c not too big in magnitude. Thus this probability can depend on x' and x'' only through the length $x'' - x'$: i.e., $\mathbf{P}\{\xi \in [x', x'']\} = f(x'' - x')$. Furthermore, using the idea of mutually exclusive events, the probability of falling in a subinterval of length $l + l'$ is equal to the sum of probabilities of falling in subintervals of length l and length l'. So $f(l+l') = f(l) + f(l')$ for any allowed l, l'. It can be shown that such a function is either linear $f(l) \propto l$ or unbounded for every l. However, $f(l) \leq f(b-a) = 1$ for every $l \leq b-a$. It follows that $f(l) = l/(b-a)$. Thus

$$\mathbf{P}\{x' \leq \xi \leq x''\} = \frac{x''-x'}{b-a} = \int_{x'}^{x''} \frac{dx}{b-a}. \tag{34}$$

Thus ξ is a continuous random variable with probability density

$$p_\xi(x) = \begin{cases} 1/(b-a) & \text{if} \quad a \leq x \leq b \\ 0 & \text{if} \quad x < a \quad \text{or} \quad x > b. \end{cases} \tag{35}$$

It is said to have a uniform distribution.

• **Statistical characterization of a probability distribution.** Since random variables take a variety of values in different trials of an experiment, we say that a random variable fluctuates. In physics such random variations (say in the position of a particle or pressure of a gas) typically arise due to quantum and thermal fluctuations. This means we need to treat the behavior of random variables probabilistically. We can characterize the distribution of a random variable using some statistical quantities. The mean value is the simplest of them. Fluctuations around the mean measure the width of the probability density function and are encoded in quantities such as the variance. Moments are more general quantities that measure fluctuations. In what follows, we will introduce these quantities and study their properties and interpretation.

6 Expectation or mean value

Expectation value. The expectation value or mean/average value of a discrete random variable with probability distribution $P_\xi(x) = \mathbf{P}\{\xi = x\}$ is defined as a weighted sum of all values that the random variable takes:

$$\mathbf{E}\xi = \langle \xi \rangle = \sum_{i \in I} x_i \mathbf{P}\{\xi = x_i\} = \sum_{i \in I} x_i P_\xi(x_i), \tag{36}$$

assuming the series converges absolutely. If $\eta = \varphi(\xi)$ is some function of the random variable ξ, then

$$\mathbf{E}\eta = \langle \varphi(\xi) \rangle = \sum_i \varphi(x_i) P_\xi(x_i). \tag{37}$$

To see why, we note that η is a discrete random variable that takes the values $y = \varphi(x)$. If φ is not one-to-one there may be several values of x corresponding to a given value of y. Thus,

$$P_\eta(y) = \mathbf{P}\{\eta = y\} = \sum_{x:\varphi(x)=y} P_\xi(x). \tag{38}$$

Consequently,

$$\langle \eta \rangle = \sum_y y P_\eta(y) = \sum_y y \sum_{x:\varphi(x)=y} P_\xi(x) = \sum_i \varphi(x_i) P_\xi(x_i). \tag{39}$$

• The expected value of a continuous random variable ξ with probability density function $p_\xi(x)$ is

$$\langle \xi \rangle = \int_{-\infty}^{\infty} x p_\xi(x)\, dx. \tag{40}$$

As before, if $\eta = \varphi(\xi)$, then

$$\langle \varphi(\xi) \rangle = \int_{-\infty}^{\infty} \varphi(x) p_\xi(x)\, dx. \tag{41}$$

• More generally, if $\varphi(\xi, \eta)$ is a function of a pair of discrete or continuous random variables with probability distribution $P_{\xi,\eta}(x, y)$ or probability density function $p_{\xi,\eta}(x, y)$, then

$$\begin{aligned} \langle \varphi(\xi, \eta) \rangle &= \sum_{i,j} \varphi(x_i, y_j) P_{\xi,\eta}(x_i, y_j) \quad \text{or} \\ \langle \varphi(\xi, \eta) \rangle &= \iint \varphi(x, y) P_{\xi,\eta}(x, y)\, dxdy \end{aligned} \tag{42}$$

Properties of expectation value. We list some basic properties of the expectation value of discrete as well as continuous random variables

1. $\langle 1 \rangle = 1$. One way to interpret this is via $\langle 1 \rangle = \int_{-\infty}^{\infty} p_\xi(x) dx = 1$.
2. $\langle c\xi \rangle = c\langle \xi \rangle$ for any real constant c.
3. $\langle \xi_1 + \xi_2 \rangle = \langle \xi_1 \rangle + \langle \xi_2 \rangle$ for a pair of random variables ξ_1 and ξ_2 with expectation values appearing on the right.
4. The expectation of a nonnegative random variable is nonnegative: If $\xi \geq 0$, then $\langle \xi \rangle \geq 0$ and more generally, if $\xi_1 \leq \xi_2$, then $\langle \xi_1 \rangle \leq \langle \xi_2 \rangle$.
5. Suppose ξ_1 and ξ_2 are independent random variables. Since the joint probability density of a pair of independent random variables factorizes, $\langle \xi_1 \xi_2 \rangle = \langle \xi_1 \rangle \langle \xi_2 \rangle$.

• The expectation value of a random variable ξ that is **uniformly distributed** in the interval $[a, b]$ is $(a + b)/2$. In fact, the probability density is $1/(b - a)$ for $a \leq x \leq b$ and zero outside the interval, so that

$$\langle \xi \rangle = \int_a^b \frac{x\, dx}{b - a} = \frac{a + b}{2}. \tag{43}$$

7 Mean square, Chebyshev's inequality and variance

Mean square value. By the mean square value of a real random variable ξ we mean the expectation value of ξ^2:

$$\langle \xi^2 \rangle = \sum_i x_i^2 P_\xi(x_i) \quad \text{or} \quad \langle \xi^2 \rangle = \int x^2 p_\xi(x)\, dx \tag{44}$$

according as ξ is discrete or continuous.

Chebyshev's inequality. For any real random variable and any $\epsilon > 0$,

$$\mathbf{P}\{|\xi| > \epsilon\} \leq \frac{1}{\epsilon^2} \langle \xi^2 \rangle. \tag{45}$$

We will use Chebyshev's inequality to establish the (weak) law of large numbers in §**??**. To establish Chebyshev's inequality, consider the new 'piece-wise constant' random variable η defined for any $\epsilon > 0$ by

$$\eta = \begin{cases} 0 & \text{if} \quad \xi^2 \leq \epsilon^2 \\ \epsilon^2 & \text{if} \quad \xi^2 > \epsilon^2 \end{cases} \tag{46}$$

Roughly, where ξ^2 is smaller than an arbitrary threshold value ϵ^2, η vanishes while when when ξ^2 exceeds the threshold, η takes the constant threshold value. By construction, $\eta \leq \xi^2$. It follows that

$$\langle \eta \rangle \leq \langle \xi^2 \rangle \quad \text{or} \quad \epsilon^2 \mathbf{P}\{|\xi| > \epsilon\} \leq \langle \xi^2 \rangle, \tag{47}$$

which is Chebyshev's inequality (45). To heuristically interpret Chebyshev's inequality, suppose $\frac{1}{\epsilon^2}\langle \xi^2 \rangle < \delta$. Then $\mathbf{P}\{|\xi| \leq \epsilon\} \geq 1 - \delta$. So if δ is small, then $|\xi| \leq \epsilon$ with a high probability. In particular, if the mean square value $\langle \xi^2 \rangle = 0$, then $\xi = 0$ with probability one. Roughly speaking, if the mean square value is small, then the random variable is likely to be small.

Variance or dispersion of a random variable. By the variance or dispersion of the random variable ξ we mean the mean square value of $\xi - \langle \xi \rangle$:

$$\mathrm{var}(\xi) = \mathbf{D}\xi = \langle (\xi - \langle \xi \rangle)^2 \rangle = \langle (\xi^2 - 2\xi\langle \xi \rangle + \langle \xi \rangle^2) \rangle = \langle \xi^2 \rangle - \langle \xi \rangle^2. \tag{48}$$

Note that $\langle \xi - \langle \xi \rangle \rangle$ is identically zero. This is why we squared it before taking the expectation value.

• **Standard deviation.** The square-root of the variance is called the standard deviation $\sigma(\xi)$

$$\sigma(\xi) = \sqrt{\mathrm{var}(\xi)}. \tag{49}$$

The dispersion and standard deviation are measures of the fluctuations of ξ around its mean value.

• Properties of the dispersion or variance

1. $\mathrm{var}\ (1) = 0$
2. $\mathrm{var}\ (c\xi) = c^2\ \mathrm{var}\ (\xi)$ for any real number c.
3. If ξ_1 and ξ_2 are independent random variables, then

$$\mathrm{var}\ (\xi_1 + \xi_2) = \ \mathrm{var}\ (\xi_1) + \ \mathrm{var}\ (\xi_2). \tag{50}$$

   Show that this is true using the property that the expectation value of a product of independent random variables factorizes: $\langle \xi_1 \xi_2 \rangle = \langle \xi_2 \rangle \langle \xi_2 \rangle$.

• What is the variance of a random variable ξ with a uniform distribution on the interval $[a, b]$?

Covariance. Given a pair of random variables ξ and η, we define their covariance by

$$\mathrm{cov}(\xi, \eta) = \langle (\xi - \langle \xi \rangle)(\eta - \langle \eta \rangle) \rangle. \tag{51}$$

Evidently, it is symmetric: $\mathrm{cov}(\xi, \eta) = \mathrm{cov}(\eta, \xi)$. The covariance is a measure of the correlation between random variables. If ξ and η are independent random variables then their covariance vanishes. On the other hand, if $\eta = \xi$ then $\mathrm{cov}(\xi, \eta) = \mathrm{cov}(\xi, \xi) = \mathrm{var}\ \xi$.

• For any pair of real random variables (not necessarily independent), show that

$$\mathrm{var}\ (\xi_1 + \xi_2) = \ \mathrm{var}\ (\xi_1) + \ \mathrm{var}\ (\xi_2) + 2\,\mathrm{cov}(\xi_1, \xi_2). \tag{52}$$

8 Moments, cumulants, generating and characteristic functions

Moments. Given a random variable ξ, its moments G_n, when they exist, are defined as

$$G_n = \mathbf{E}\xi^n = \langle \xi^n \rangle \quad \text{for} \quad n = 0, 1, 2, \ldots. \tag{53}$$

Evidently, $G_0 = 1$, $G_1 = \mathbf{E}\xi$ is the expected value and $G_2 = \langle \xi^2 \rangle$ is the mean square value. Moreover, the variance is given by $\mathbf{D}\xi = G_2 - G_1^2$. If ξ is discrete, taking the values x_i with probability $P_\xi(x_i)$, then

$$G_n = \mathbf{E}\xi^n = \sum_i P_\xi(x_i) x_i^n \tag{54}$$

while for a continuous random variable with probability density $p_\xi(x)$,

$$G_n = \int_{-\infty}^{\infty} x^n p_\xi(x)\, dx. \tag{55}$$

Of course, not all moments may exist. If $p_\xi(x)$ goes to zero exponentially fast as $|x| \to \infty$, then all moments are guaranteed to exist.

Generating function for a discrete random variable. Suppose ξ is a discrete random variable taking the values $0, 1, 2, \ldots$ with probabilities $\mathbf{P}\{\xi = k\} = P_\xi(k)$ for $k = 0, 1, 2, 3, \ldots$. Associated to such a discrete random variable is a generating function, which is the function of a complex variable defined as

$$F_\xi(z) = \sum_{k=0}^{\infty} P_\xi(k) z^k \quad \text{for} \quad |z| \le 1. \tag{56}$$

Clearly, $F_\xi(1) = 1$. Since P_ξ is a probability distribution, the series converges for $|z| = 1$. In fact, the series defines an analytic function for $|z| < 1$. What is more, we may recover the probability distribution of ξ via derivatives of F_ξ at $z = 0$. In fact,

$$P_\xi(k) = \frac{1}{k!} F_\xi^{(k)}(0). \tag{57}$$

Interestingly, for any fixed z, $F_\xi(z)$ may be interpreted as the expectation value of the random variable z^ξ:

$$F_\xi(z) = \langle z^\xi \rangle = \sum_{k \ge 0} P_\xi(k) z^k. \tag{58}$$

Differentiating this expression successively with respect to z and putting $z = 1$ allows us to express the moments of ξ in terms of derivatives of F at $z = 1$. For instance,

$$F'(z) = \langle \xi z^{\xi - 1} \rangle \quad \Rightarrow \quad F'(1) = \langle \xi \rangle = G_1, \tag{59}$$

and

$$F''(z) = \langle \xi(\xi - 1) z^{\xi - 2} \rangle \quad \Rightarrow \quad G_2 = \langle \xi^2 \rangle = F''(1) + F'(1). \tag{60}$$

• The generating function of a sum of independent random variables each taking the values $k = 0, 1, 2, 3, \ldots$ is the product of individual generating functions. In fact, suppose $\xi = \xi_1 + \cdots + \xi_n$. Then

$$F_\xi(z) = \langle z^{\xi_1 + \cdots \xi_n} \rangle = \langle z^{\xi_1} z^{\xi_2} \cdots z^{\xi_n} \rangle = \prod_{i=1}^{n} \langle z^{\xi_i} \rangle = \prod_{i=1}^{n} F_{\xi_i}(z). \tag{61}$$

A similar factorization is used in calculating the partition function of a system of noninteracting (free) particles in statistical mechanics.

Characteristic function. Given a real random variable ξ, its characteristic function $f_\xi(t)$ is defined as

$$f_\xi(t) = \mathbf{E} e^{i\xi t} \quad \text{for} \quad t \in \mathbb{R}. \tag{62}$$

• For a **discrete random variable** that takes the values $k = 0, 1, 2, \ldots$, we see that the characteristic function reduces to the generating function evaluated on the boundary of the unit circle ($|z| = 1$) in the complex plane:

$$f_\xi(t) = F_\xi(z = e^{it}) = \sum_k P_\xi(k) e^{ikt}. \tag{63}$$

In fact, in this case, the characteristic function is a Fourier series with Fourier coefficients given by the probabilities $P_\xi(k)$.

• For a **continuous random variable** ξ, the characteristic function is the Fourier transform of the probability density function $p_\xi(x)$:

$$f_\xi(t) = \langle e^{i\xi t} \rangle = \int_{-\infty}^{\infty} p_\xi(x) e^{ixt} dx. \tag{64}$$

By inverting the Fourier transform, we may recover the probability density function (where it is well-behaved) from the characteristic function

$$p_\xi(x) = \int_{-\infty}^{\infty} f_\xi(t) e^{-ixt} \frac{dt}{2\pi}. \tag{65}$$

Moments from characteristic function. Provided the moments exist, by differentiating under the integral sign, we may obtain the moments as successive derivatives of the characteristic function evaluated at $t = 0$. To begin with, $f(0) = \langle 1 \rangle = G_0 = 1$. Next,

$$f'(t) = i \int x e^{ixt} p_\xi(x)\, dx \quad \Rightarrow \quad f'(0) = i\langle \xi \rangle = iG_1. \tag{66}$$

Similarly, $f''(0) = i^2 G_2$ and more generally,

$$G_n = (-i)^n f^{(n)}(0) \quad \text{for} \quad n = 0, 1, 2, \ldots. \tag{67}$$

All moments need not exist. As long as $\langle |\xi|^k \rangle$ exists, the above formulae for G_n hold for $n < k$.

Characteristic function as generating series for moments. By expanding $e^{i\xi t}$ in a power series and assuming the probability density is sufficiently well-behaved to permit interchanging the order of integration and summation, we may express the characteristic function as a power series with coefficients proportional to the moments:

$$f_\xi(t) = \int \sum_0^\infty \frac{1}{n!} t^n (ix)^n p_\xi(x)\, dx = \sum_{n=0}^\infty \frac{(it)^n}{n!} G_n. \tag{68}$$

Thus, we may view the characteristic function as a generating function (or series) for moments. Evidently, if the characteristic function is analytic at $t = 0$, then the generating series of moments converges. This happens if the moments do not grow too fast in magnitude (e.g., not faster than exponentially, i.e., $|G_n| \le c^n$ for some constant $c > 0$).

• What is the characteristic function $f_\xi(t) = \langle e^{i\xi t} \rangle$ of the uniform distribution on the unit interval $[0, 1]$?

• Closely related to the characteristic function is the moment generating function, defined as $M_\xi(t) = \langle e^{t\xi}\rangle$ for real t. Crudely, it is the characteristic function evaluated at imaginary arguments. Unlike the characteristic function which is the expectation value of the bounded random variable $e^{i\xi t}$ the moment generating function is the expectation value of the unbounded random variable $e^{t\xi}$. Thus, the latter may fail to exist if the probability density does not vanish sufficiently fast for large $|\xi|$.

• **Cumulants.** The cumulants $C_{n\geq 0}$ of a real-valued probability distribution provide an alternative to its moments G_n. They may be defined as the coefficients in the series expansion of the logarithm of the characteristic function

$$W_\xi(t) = \log f_\xi(t) = \log\langle e^{i\xi t}\rangle = \sum_{n=0}^{\infty} \frac{(it)^n}{n!} C_n. \tag{69}$$

The first few cumulants are

$$C_0 = 0, \quad C_1 = \langle\xi\rangle = G_1, \quad C_2 = \text{ var}\,\xi = G_2 - G_1^2, \quad C_3 = \langle(\xi - \langle\xi\rangle)^3\rangle. \tag{70}$$

Derive formulae for C_n in terms of G_k for $n = 0, 1, 2, 3$.

• We will see later that the third and higher order cumulants of a Gaussian vanish.

9 Bernoulli trials and the binomial distribution

Bernoulli trials are identical independent experiments in each of which an event A may occur with probability p and fail to occur with probability $q = 1 - p$. Occurrence of A is called a success and its nonoccurrence a failure. It is convenient to introduce the 'Bernoulli' random variable ξ_k associated to the k^{th} Bernoulli trial, taking the values 1 or 0 depending on whether the trial is a success or a failure. With this understanding, each elementary event ω in n consecutive Bernoulli trials may be described by an n digit binary number such as $0100111\cdots01011$.

Binomial distribution. Let us define the random variable $\xi = \xi_1 + \cdots + \xi_n$, which is the number of successes in n Bernoulli trials. We wish to find the probability of k successes i.e., $\mathbf{P}\{\xi = k\}$. Since the trials are independent, the probability of any elementary event ω with k successes and $n-k$ failures is $\mathbf{P}(\omega) = p^k q^{n-k}$. There are $\binom{n}{k}$ elementary events ω with k successes. Thus $\mathbf{P}\{\xi = k\} = \binom{n}{k} p^k q^{n-k}$. This leads us to the binomial distribution

$$P_\xi(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for} \quad k = 0, 1, 2, \cdots, n. \tag{71}$$

It is the probability distribution of the discrete random variable ξ equal to the number of successes in n Bernoulli trials with p being the probability of success in each trial.

Mean and variance of Binomial random variable ξ. To find these it is not necessary to evaluate $\sum_k P_\xi(k)k$ and $\sum_k P_\xi(k)k^2$ etc. Instead we note that $\xi = \xi_1 + \cdots + \xi_n$ implies that $\langle\xi\rangle = \langle\xi_1\rangle + \cdots + \langle\xi_n\rangle$ and since ξ_i are independent, var $(\xi) =$ var $(\xi_1) + \cdots +$ var (ξ_n). What is more, since ξ_k are identically distributed for $i = 1, 2, 3, \cdots, n$, $\langle\xi_1\rangle = \cdots = \langle\xi_n\rangle$ and var $\xi_1 =$ var $\xi_2 = \cdots =$ var ξ_n. Furthermore, $\langle\xi_k\rangle = p \cdot 1 + (1-p) \cdot 0 = p$. Moreover, $\xi_k^2 = \xi_k$, so var $\xi_k = \langle\xi_k^2\rangle - \langle\xi_k\rangle^2 = \langle\xi_k\rangle - p^2 = p - p^2 = pq$. Consequently, $\langle\xi\rangle = np$ and var$(\xi) = npq$. Thus there are on average np successes in n Bernoulli trials with a variance of $np(1-p)$.

• Show that the generating function of a binomial random variable with parameters n, p is

$$F_\xi(z) = (pz + q)^n. \tag{72}$$

10 Poisson distribution: a limit of the binomial distribution

Poisson distribution. The Poisson distribution is a limit of the binomial distribution as the number of trials $n \to \infty$ and the probability of success $p \to 0$ while the mean number of successes $np = a$ has a finite limit (we will apply this to radioactive decay shortly). In fact,

$$\binom{n}{k} p^k (1-p)^{n-k} \to \frac{a^k}{k!} e^{-a} \quad \text{for} \quad k = 0, 1, 2, \ldots. \tag{73}$$

To see this we consider $k = 0, 1, 2, \ldots$ successively. To begin with,

$$P_\xi(0) = (1-p)^n = (1 - a/n)^n \to e^{-a}. \tag{74}$$

Furthermore,

$$\frac{P_\xi(k)}{P_\xi(k-1)} = \frac{n! p^k q^{n-k}}{(n-k)! k!} \frac{(n-k+1)!(k-1)!}{n! p^{k-1} q^{n-k+1}} = \frac{p(n-k+1)}{kq} \to \frac{a}{k} \tag{75}$$

as $n \to \infty$ and $p \to 0$. It follows that

$$\begin{aligned} P_\xi(1) &= \frac{a}{1} P_\xi(0) = a e^{-a}, \quad P_\xi(2) = \frac{a}{2} P_\xi(1) = \frac{a^2}{1 \cdot 2} e^{-a}, \\ P_\xi(3) &= \frac{a}{3} P_\xi(2) = \frac{a^3}{1 \cdot 2 \cdot 3} e^{-a}, \quad \ldots, P_\xi(k) = \frac{a}{k} P_\xi(k-1) = \frac{a^k}{k!} e^{-a}. \end{aligned} \tag{76}$$

• Thus, the probability of k successes in n Bernoulli trials when the probability p of each success is small and the number of trials $n \to \infty$ holding $a = np$ fixed is

$$P_\xi(k) = \frac{a^k}{k!} e^{-a} \quad \text{for} \quad k = 0, 1, 2, \ldots. \tag{77}$$

A random variable ξ taking values $k = 0, 1, 2, \ldots$ and possessing this distribution is said to have a Poisson distribution with parameter a. It is straightforward to check that the distribution is normalized: $\sum_{k \geq 0} P_\xi(k) = e^a e^{-a} = 1$.

• Unlike the binomial distribution that depends on two parameters n and p, the Poisson distribution depends only on one positive real parameter $a > 0$, which is equal to its mean value. This is immediate since the mean value of a Binomial random variable is np, which in the limit considered is equal to a. Thus,

$$\langle \xi \rangle = \sum_{k=0}^{\infty} k P_\xi(k) = a. \tag{78}$$

• The variance of a binomial random variable is npq which becomes a in the limit $n \to \infty$, $p \to 0$ and $q \to 1$. Thus, the variance of a Poisson distribution is the same as its mean, a.

• The **generating function** of the Poisson distribution with mean a is

$$F_\xi(z) = \sum_{k=0}^{\infty} P_\xi(k) z^k = \sum_{k=0}^{\infty} \frac{(az)^k}{k!} e^{-a} = e^{a(z-1)}. \tag{79}$$

Modeling radioactive decay via a Poisson distribution. A gram of radium has about 10^{22} atoms. It gradually decays to radon through the emission of about 10^{10} alpha particles per second. Suppose there are a large number n_0 of radium atoms in a container at $t = 0$. The atoms are sufficiently far separated to justify the assumption that each atom decays independently of all the others. Moreover, since they are identical, each radium atom has the same probability $p(t)$ to decay in t seconds. In fact, for moderate times, this decay probability is quite small, $p(1) \approx 10^{10-22} = 10^{-12}$. Thus, it is natural to model the radioactive decay of radium atoms in terms of a large number of Bernoulli trials, each with a small probability of success (decay). Let the random variable $\xi(t)$ denote the number of alpha particles emitted in t seconds. It is the number of successes in n_0 Bernoulli trials with probability of success $p(t)$. Since n_0 and $p(t)$ are small, the binomial distribution of $\xi(t)$ may be well approximated by a Poisson distribution

$$\mathbf{P}\{\xi(t) = k\} = \frac{a^k}{k!}e^{-a} \quad \text{for} \quad k = 0, 1, 2, \ldots \quad \text{where} \quad a = \langle \xi(t) \rangle = n_0 p(t) \tag{80}$$

is the average number of alpha particles emitted in t seconds. This probability distribution agrees well with experimental measurements of the number of alpha particles emitted in t seconds.

Related continuous probability distributions. Interestingly, if we think of $k = 0, 1, 2, \ldots$, as a parameter, then

$$p_\eta^{(k)}(s) = \frac{s^k e^{-s}}{k!} \quad \text{for} \quad s \geq 0 \tag{81}$$

may be viewed as a probability density function for a continuous positive real random variable η. We verify that

$$\int_0^\infty p_\eta^{(k)}(s) ds = \int_0^\infty \frac{s^k e^{-s}}{k!} = 1 \quad \text{for any} \quad k = 0, 1, 2, \ldots. \tag{82}$$

Interpretation: The probability density $p_\eta^{(k)}(s)$ arises in the spectral statistics of (unfolded) quantum energy levels of a classically integrable system. It turns out that $p_\eta^{(k)}(s)ds$ is the probability that the spacing between k^{th} nearest neighbor energy levels lies between s and $s + ds$.

11 Gaussian or Normal distribution

De Moivre-Laplace limit theorem. Suppose $\xi_1, \xi_2, \ldots, \xi_n$ are n independent identically distributed ('iid') random variables, each taking the values 1 and 0 with probabilities p and $q = 1 - p$. ξ_k are of course the 'Bernoulli' random variables introduced in the context of Bernoulli trials. As before, we define the sum

$$S_n = \xi_1 + \cdots + \xi_n \tag{83}$$

which is a random variable (previously denoted ξ) taking the values $0, 1, \ldots, n$ with mean and variance $\langle S_n \rangle = np$ and $\text{var}(S_n) = npq$. We know that S_n is the number of successes in n Bernoulli trials. It has a binomial distribution

$$\mathbf{P}\{S_n = k\} = \binom{n}{k} p^k q^{n-k}. \tag{84}$$

Now consider the normalized sum

$$S_n^* = \frac{S_n - \langle S_n \rangle}{\sqrt{\mathrm{var}(S_n)}}, \tag{85}$$

which is a random variable taking the values $x = (k - np)/\sqrt{npq}$ for $k = 0, 1, 2, \ldots, n$ with probabilities given by the binomial formula $\binom{n}{k} p^k q^{n-k}$. Now, it can be shown using the Stirling approximation ($n! \sim \sqrt{2\pi n} n^n e^{-n}$) that as $n \to \infty$, S_n^* tends to a continuous real random variable with probability distribution given by

$$\lim_{n \to \infty} \mathbf{P}\{x' \le S_n^* \le x''\} = \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-x^2/2}\, dx. \tag{86}$$

This result was discovered by de Moivre in 1733.

Gaussian probability density. The corresponding limiting probability density

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for} \quad -\infty < x < \infty \tag{87}$$

is called the (standard) Gaussian or normal distribution. The graph of the probability density is a bell-shaped curve. It is an even function.

- The Gaussian probability density occurs in the work of de Moivre from 1733. Laplace considered normal random variables around 1780. They are named after Gauss, who discussed them in 1809.
- A random variable ξ with the standard Gaussian probability density is called a standard Gaussian random variable. It has mean zero and variance one. In fact,

$$\langle \xi \rangle = \int x p(x)\, dx = 0 \tag{88}$$

since the integrand is odd. On the other hand,

$$\mathrm{var}(\xi) = \sigma_\xi^2 = \langle \xi^2 \rangle - \langle \xi \rangle^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx = 1. \tag{89}$$

Normal cumulative distribution function. The corresponding cumulative distribution function is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du. \tag{90}$$

We verify that $\Phi(x) \to 0, 1$ as $x \to \mp\infty$ as required of a cumulative distribution function. Since $p(x) = p(-x)$, it follows that $\Phi(-x) = 1 - \Phi(x)$. In particular $\Phi(0) = 1/2$. Since $\Phi(x)$ is the probability that $\xi \le x$,

$$\mathbf{P}\{|\xi| \le x\} = 2(\Phi(x) - \Phi(0)). \tag{91}$$

Normal distribution with mean a and variance σ^2. Note that if X is a standard Gaussian random variable with mean zero and variance one, then $Y = \sigma(X + a)$ is a normal random variable with mean a and variance σ^2. To obtain the pdf $p_Y(y)$ of Y from the standard gaussian for X we change variables $y = \sigma(x + a)$ and $dy = \sigma dx$ in $p_X(x)dx$ and use

$$p_Y(y)dy = p_X(x)dx \tag{92}$$

to deduce that

$$p_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-a)^2/2\sigma^2} \tag{93}$$

a random variable with this probability density function is called a normal random variable with mean a and variance σ^2.

Error function. The error function is conventionally defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt. \tag{94}$$

It is closely related to the cumulative distribution function of a standard gaussian random variable:

$$\text{erf}(x) = 2(\Phi(\sqrt{2}x) - \Phi(0)) = 2\Phi(\sqrt{2}x) - 1. \tag{95}$$

From this we deduce that $\text{erf}(x)$ is the probability that a standard Gaussian random variable ξ is at most $\sqrt{2}x$ in magnitude:

$$\text{erf}(x) = \mathbf{P}\{|\xi| \le \sqrt{2}x\}. \tag{96}$$

• Suppose ξ is a normal random variable with mean a and variance σ^2. Then the probability $\mathbf{P}\{|\xi - a| < n\sigma\}$ that ξ lies within $n\sigma$ of μ for $n = 1, 2, 3, \cdots$ are $\approx 0.683, 0.954, 0.997$. So with $\approx 99.7\%$ probability, a gaussian random variable takes values within 3σ of its mean. It lies within one standard deviation of the mean with probability $\approx 68\%$.

Characteristic function of the Gaussian. Suppose ξ is a standard Gaussian random variable with probability density $e^{-x^2/2}/\sqrt{2\pi}$. It is possible to show that its characteristic function is also a Gaussian

$$f_\xi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{ixt}\, dx = e^{-t^2/2}. \tag{97}$$

To see this, we complete the square to express this as a Gaussian integral:

$$e^{-x^2/2} e^{ixt} = e^{-\frac{1}{2}(x^2 - 2ixt)} = e^{-\frac{1}{2}\{(x-it)^2 + t^2\}}. \tag{98}$$

Thus

$$f_\xi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\{(x-it)^2+t^2\}}\, dx = e^{-t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = e^{-t^2/2} \tag{99}$$

where we put $y = x - it$ and used $dy = dx$. It follows that the cumulant generating function of the standard Gaussian is

$$W_\xi(t) = \log f_\xi(t) = -t^2/2. \tag{100}$$

Consequently, the cumulants ($C_n = (-i)^n W^{(n)}(0)$) of the standard Gaussian are $C_0 = 0, C_1 = 0, C_2 = 1$ and $C_n = 0$ for $n \geq 3$. The standard gaussian may be characterized as the distribution for which all cumulants other than the second vanish and $C_2 = 1$.