3 Probability Models

As we discussed in Chapter 2, Bayes' theorem and the Bayesian approach to probability constitute central components of the scientific method. Recall that Bayes' theorem may be used, in particular, to infer the posterior probability of a scientific hypothesis, H_i based on measured data D according to $p(H_i|D, I) = p(D|H_i, I)p(H_i|I)/p(D|I)$. Additional hypotheses, H_j, with $j \neq i$, may also be tested in the same way. The posterior probabilities $p(H_i|D, I)$ associated with these hypotheses may then be compared to establish which is best supported by the measured data D. Calculations of posterior probabilities, however, require one has techniques to calculate both the likelihoods $p(D|H_i, I)$ and the priors $p(H_i|I)$. The determination of prior probabilities is a topic all of its own that we address in Chapter 7. In this chapter, we discuss, through several concrete examples, how prior knowledge about a phenomenon or system, denoted I, may be formulated to obtain **data probability models**, that is, probability distributions and probability densities that model phenomena of interest and can thus be used to calculate, based on specific hypotheses, the likelihood of data. We begin in §3.1, §3.2, and §3.3 with basic models including the binomial distribution, the multinomial distribution, and the Poisson distribution. Several continuous distributions, including the uniform distribution, the exponential distribution, and the normal distribution, are introduced in later sections.

Professional scientists make use of a vast array of functions and probability distributions that cannot all be covered in a single text. The functions discussed in this chapter were selected for their foundational value, their broad range of application, or because they are needed in later chapters. Short examples of application are included in this chapter whenever possible and more extensive or specialized uses are discussed in later chapters of the book. Clearly, this chapter may be read both as an introduction to practical probability distributions/densities as well as a reference manual for such distributions.

3.1 Bernoulli and Binomial Distributions

3.1.1 Definition of the Bernoulli Distribution

Named after Swiss mathematician Jacob Bernoulli[1] (1654–1705), the **Bernoulli distribution** is a discrete probability distribution that takes exclusively two values: 1 with success

[1] Contributed works in calculus, particularly the calculus of variations, and in the field of probability.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

probability ε, and 0 with "failure" probability $1 - \varepsilon$:

$$p_{\text{Ber}}(n|\varepsilon) = \begin{matrix} 1-\varepsilon & \text{probability of failure, } n = 0, \\ \varepsilon & \text{probability of success, } n = 1. \end{matrix} \tag{3.1}$$

It may alternatively be written

$$p_{\text{Ber}}(n|\varepsilon) = (1-\varepsilon)^{1-n}\varepsilon^n, \tag{3.2}$$

where $n = 1$ stands for success and $n = 0$ for failure.

Arguably the simplest example of the Bernoulli distribution is the single toss of a fair coin, which has a probability $\varepsilon = 0.5$ to come up heads or tails. The particle detection process may also be regarded as a "Bernoulli experiment." In this case, ε represents the probability of detecting a single particle with a given apparatus.

3.1.2 Definition of the Binomial Distribution

In the context of the Bayesian approach to probability (i.e., probability as logic), one can formulate prior information about the system, denoted I, by stating that the single toss of a coin (or an attempt to detect a particle) is an event, E, with two possible outcomes that we denote S, "*the event was a success*," and $\overline{S}$, "*the event was NOT a success*." The event E may then be written

$$E = S + \overline{S}, \tag{3.3}$$

where the $+$ operator corresponds to a logical OR. Since S and $\overline{S}$ are mutually exclusive (i.e., $S, \overline{S}$ = false, or $S \cap \overline{S} = 0$), the probability of the event E may be written

$$p(E|I) = p(S|I) + p(\overline{S}|I) = \varepsilon + (1-\varepsilon) = 1. \tag{3.4}$$

A series of n tosses may then be regarded as a succession of Bernoulli events, each with two possible outcomes, but a unit probability. This then leads us naturally to the introduction of the binomial distribution.

The **binomial distribution** arises in the context of identical, independent measurements, tosses, or samplings, repeated a finite number of times. As such, it is an extension of the Bernoulli distribution for $n > 1$. Each measurement, observation, or sample is assumed to be an independent trial with a probability of success denoted by ε. A set of N measurements can also be regarded as a single measurement, or event, yielding n successful trials and $N - n$ "failures." If one were to carry out the N trial measurements several times, the resulting value of n would occur with a probability given by the binomial distribution, $p_B(n|N, p)$:

$$p_{\text{B}}(n|N,p) = \frac{N!}{n!(N-n)!}\varepsilon^N(1-\varepsilon)^{N-n}. \tag{3.5}$$

The binomial distribution may evidently be derived using the frequentist approach to probability. It is more interesting to discover, however, how it naturally arises also in the context of the Bayesian approach, and in particular, how the calculus of predicates of probability as logic provides a programmatic and robust recipe to derive this result.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Prior information about the series of tosses, I, may be summarized in two statements: each toss is a Bernoulli events with two mutually exclusive outcomes, and the n tosses are mutually independent, i.e., the outcome of one does not affect the outcome of the others.

As stated earlier, a single toss (or draw) may be regarded as an event, $E = S + \overline{S}$, with unit probability. A series of n tosses may also be regarded as an event, which we shall denote B_n. This event consists of n independent Bernoulli tosses we note E_i. Since the occurrence of these n mutually independent events is known, the aggregate event B_n may then be viewed as an AND between all these individual elementary events: the first event happened and then the second, and then the third, and so forth. One may then write

$$B_n \equiv E_1, E_2, \ldots, E_n \tag{3.6}$$

Additionally, since each event E_i has two possible outcomes, S_i and $\overline{S}_i$, one can also write

$$B_n = (S_1 + \overline{S}_1), (S_2 + \overline{S}_2), \ldots, (S_n + \overline{S}_n). \tag{3.7}$$

Recall that the commas represent logical ANDs between the different events. The AND may then be carried out as a multiplication of the n operands of the expression. For $n = 2$, one gets

$$B_2 = (S_1 + \overline{S}_1), (S_2 + \overline{S}_2) \tag{3.8}$$

$$= S_1, S_2 + S_1, \overline{S}_2 + \overline{S}_1, S_2 + \overline{S}_1, \overline{S}_2 \tag{3.9}$$

while for $n = 3$ one gets

$$\begin{aligned} B_3 = {} & S_1, S_2, S_3 + S_1, S_2, \overline{S}_3 + S_1, \overline{S}_2, S_3 + \overline{S}_1, S_2, S_3 \\ & + S_1, \overline{S}_2, \overline{S}_3 + \overline{S}_1, \overline{S}_2, S_3 + \overline{S}_1, S_2, \overline{S}_3 + \overline{S}_1, \overline{S}_2, \overline{S}_3 \end{aligned} \tag{3.10}$$

and similarly for larger values of n. The probability of a particular sequence, for instance, $S_1, S_2, \overline{S}_3$, is calculated with the product rule:

$$\begin{aligned} p(S_1, S_2, \overline{S}_3|I) &= p(S_1|I)p(S_2, \overline{S}_3|S_1, I) \\ &= p(S_1|I)p(S_2|S_1, I)p(\overline{S}_3|S_1, S_2, I) \end{aligned} \tag{3.11}$$

But recall that the prior information includes the notion that the outcomes of distinct tosses or draws are independent of one another. This means that $p(S_2|S_1, I) = p(S_2|I)$ and $p(\overline{S}_3|S_1, S_2, I) = p(\overline{S}_3|I)$. The probability of the event $S_1, S_2, \overline{S}_3$ is thus simply

$$p(S_1, S_2, \overline{S}_3) = p(S_1|I)p(S_2|I)p(\overline{S}_3|I). \tag{3.12}$$

The prior information I also tells us that the probabilities of success and failure are constant and independent of the draw. We can thus write

$$p(S_1, S_2, \overline{S}_3) = p(S|I)p(S|I)p(\overline{S}|I) = \varepsilon^2(1 - \varepsilon). \tag{3.13}$$

Similar conclusions are obtained for other combinations of S and $\overline{S}$ and for larger values of N. Indeed, we find the probability of a specific outcome depends on the number of successes and failures. For an experiment consisting of N trials, the calculation thus simply becomes a matter of keeping track of the number of ways a given number of S and $\overline{S}$ can

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

occur. This is of course given by the binomial coefficient.

$$ {}^{N}C_{n} = \frac{N!}{n!(N-n)!} \equiv \binom{N}{n}. \tag{3.14} $$

The probability of a sequence containing n times a success S and $N-n$ times a failure $\overline{S}$ is thus indeed given by the binomial distribution

$$ p(n|N,I) = \frac{N!}{n!(N-n)!} p(S|I)^{n} p(\overline{S}|I)^{(N-n)} \tag{3.15} $$

$$ = \frac{N!}{n!(N-n)!} \varepsilon^{n} (1-\varepsilon)^{(N-n)}. \tag{3.16} $$

Although the foregoing derivation is notably lengthier than a proof based on the frequentist approach, it has the important advantage of being transparent and programmatic: once the prior information I is given, the probability of any given sequence is readily written and calculated. There is no need for particular intuition or tricks.

3.1.3 Properties of the Binomial Distribution

It is relatively simple to verify that the sum of the probabilities of values of n from zero to N equals unity. Summing these different probabilities amounts to summing the probabilities of all terms of an event B_n. But an event B_n has, by construction, a probability equal to the product of the events it contains. Since all the events E_i have unit probability, the sequence event B_n thus itself also has unit probability. Alternatively, one can also use Pascal's binomial theorem, which may be written

$$ (x+y)^{N} = \sum_{n=0}^{N} \binom{N}{n} x^{N-n} y^{n}. \tag{3.17} $$

Setting $x = \varepsilon$ and $y = 1-\varepsilon$, one finds the left-hand side of Eq. (3.17) yields $(x+y)^{N} = (\varepsilon + 1 - \varepsilon)^{N} = 1$, and the right-hand side becomes the sum of the probabilities sought after.

Pascal's binomial theorem may also be used to obtain the moments of the binomial distribution. The mean, μ, of the binomial distribution is obtained by calculating the expectation value $\mathrm{E}[n]$. One gets

$$ \mu \equiv \mathrm{E}[n] = \sum_{n=0}^{N} n \frac{N!}{n!(N-n)!} p^{N} (1-p)^{N-n} = Np. \tag{3.18} $$

Similarly, the variance, σ^{2}, of the binomial distribution is given by

$$ \sigma^{2} = \mathrm{Var}[n] = \mathrm{E}[n^{2}] - (\mathrm{E}[n])^{2} = Np(1-p). \tag{3.19} $$

These two results can be verified (see Problem 3.2) using a binomial decomposition of $(1-p)^{N-n}$ in series. It is, however, simpler to calculate these, and higher moments, using the characteristic function (defined in §2.10) of the binomial distribution (see Problem 3.2).

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.1 Properties of the binomial distribution

Function	$p_B(n\|N,\varepsilon) = \frac{N!}{n!(N-n)!}\varepsilon^N(1-\varepsilon)^{N-n}$
Characteristic Fct	$\phi(t) = \left(1-\varepsilon+\varepsilon e^{it}\right)^N$
Moment Gen Fct	$M(t) = (1-\varepsilon+\varepsilon e^{t})^N$
Moments	
1st	$\mu = \mu' = N\varepsilon$
2nd	$\mu_2' = N\varepsilon(1-\varepsilon+N\varepsilon)$
3rd	$\mu_3' = Np(1-3\varepsilon+3N\varepsilon+2\varepsilon^2-3N\varepsilon^2+N^2\varepsilon^2)$
4th	$\mu_4' = N\varepsilon(1-7\varepsilon+7N\varepsilon+12\varepsilon^2-18N\varepsilon^2 + 6N^2\varepsilon^2-6\varepsilon^3+11N\varepsilon^3+N^3\varepsilon^3)$
Centered moments	
2nd	$\sigma^2 \equiv \mu_2 = N\varepsilon(1-\varepsilon)$
3rd	$\mu_3 = N\varepsilon(1-\varepsilon)(1-2\varepsilon)$
4th	$\mu_4 = N\varepsilon(1-\varepsilon)\left(3\varepsilon^2(2-N)+3\varepsilon(N-2)+1\right)$
Skewness excess	$\gamma_1 = \frac{1-2\varepsilon}{\sqrt{N\varepsilon(1-\varepsilon)}}$
Kurtosis excess	$\gamma_2 = \frac{6\varepsilon^2-6\varepsilon+1}{N\varepsilon(1-\varepsilon)}$
1st cumulant	$\kappa_1 = N\varepsilon$
Other cumulants	$\kappa_{i+1} = \varepsilon(1-\varepsilon)\frac{d\kappa_i}{d\varepsilon}$

Figure 3.1 displays examples of the binomial distribution obtained for various values of ε. The characteristic function, skewness, kurtosis, and other properties of the binomial are presented in Table 3.1.

It is interesting to consider the shape of the binomial distribution for arbitrarily large values of N, as illustrated in Figure 3.2, where the distribution is plotted as a function of

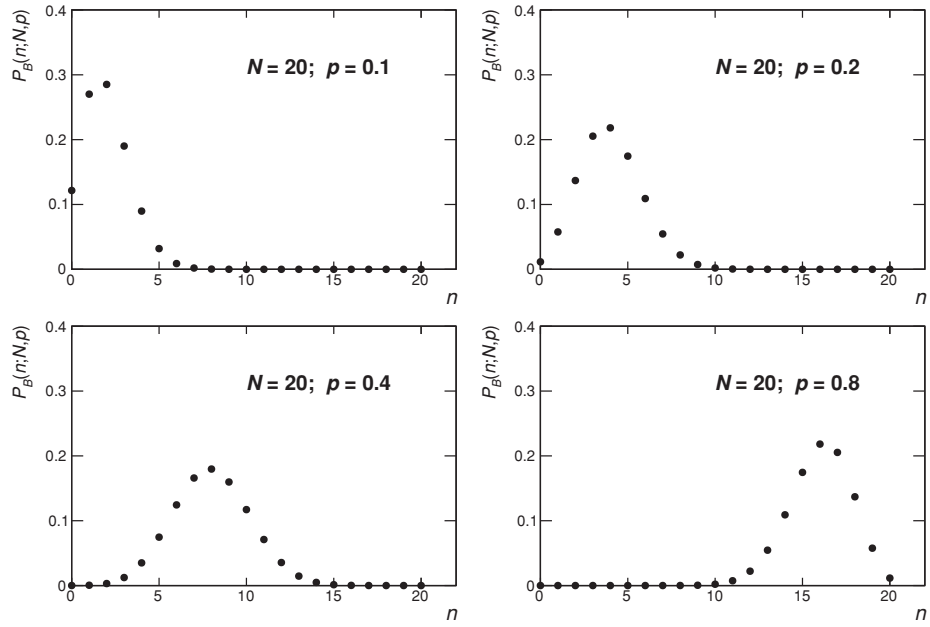


Fig. 3.1 Examples of the binomial distribution for $N = 20$ and selected values of p.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

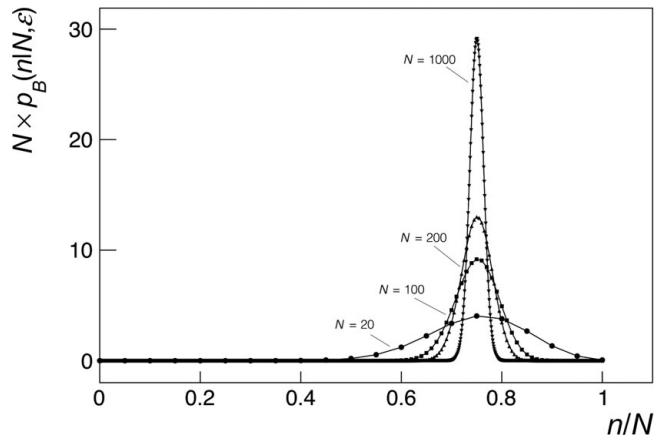


Fig. 3.2 Evolution of the binomial distribution with increasing N, plotted for $\varepsilon = 0.75$.

n/N. One finds that for increasing values of N, the distribution (plotted as a function of n/N) becomes progressively narrower, and in the limit $N \to \infty$ becomes a delta function $\delta(\varepsilon)$, that is, the distribution peaks at $n/N = \varepsilon$.

The binomial distribution finds a wide variety of applications ranging from games of chance (toss of a coin), opinion polls, and particle physics, where it plays a central role in modeling detection efficiencies, as we discuss briefly in the example provided in the next section and in greater detail in §12.2.2.

3.1.4 Example: Measurements of Particle Multiplicity Variance

Consider a measurement of particle multiplicity production in proton–proton or nucleus–nucleus collisions at relativistic energy. Assume a large acceptance detector is used to measure the particle production. For various technical reasons, it is usually impossible to measure all particles emitted within the detector acceptance. One characterizes a detector's response by stating the detector efficiency, ε, defined as the probability of measuring any given particle. Assuming the detection of a particle is not affected by the presence of other particles, we will show, in Chapter 12, that the efficiency can be estimated as the ratio of the average number of detected particles, $\langle n \rangle$, by the number of particles produced within the acceptance of the detector, $\langle N \rangle$:

$$\varepsilon = \frac{\langle n \rangle}{\langle N \rangle}. \tag{3.20}$$

The bracket notation $\langle\rangle$ is hereafter used to represent the average of the variables n and N over all measured collisions.

To obtain this result, consider that in essence, each particle has a probability ε of being detected. This implies that for a given N, the number of measured particles n should fluctuate. Given a particle is either detected or not detected (success or failure), the probability distribution function (PDF) of n is a binomial distribution with parameters N and

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

probability ε:

$$p_B(n|N,\varepsilon) = \frac{N!}{n!(N-n)!}\varepsilon^N(1-\varepsilon)^{N-n}. \tag{3.21}$$

Particle production in elementary collisions is intrinsically a stochastic phenomenon. We assume the number of produced particles, N, can be described by some a priori unknown PDF, $p_P(N)$. Its two lowest moments and variance are in principle calculated as follows:

$$\langle N\rangle = \mathrm{E}[N] = \sum p_P(N)N, \tag{3.22}$$

$$\langle N^2\rangle = \mathrm{E}[N^2] = \sum p_P(N)n^2, \tag{3.23}$$

$$\mathrm{Var}[N] = \mathrm{E}[N^2] - \mathrm{E}[N]^2. \tag{3.24}$$

But given that the PDF itself is not known a priori, these calculations are not readily possible, and one must actually *measure* these moments. The problem, of course, is that the detection efficiency is smaller than unity, and fluctuations will take place in the detection process as well as in the particle production process. For a given N, measured values n will fluctuate according to the binomial distribution. But since N itself fluctuates, one must fold the probability of production and detection. The probability of detecting n particles, $p_m(n)$, may then be expressed as

$$p_m(n) = \sum_N p_B(n|N,\varepsilon)p_P(N), \tag{3.25}$$

where the sum is taken on all possible values of N. Figure 3.3 presents plots of the distribution $p_m(n)$ for the trivial case of $p_P(N) = \delta(N)$ and illustrates how measured multiplicities can widely fluctuate for small values of efficiency. Given our interest is to estimate the moments of particle production, we proceed to calculate the moments of the measured distribution. The average measured multiplicity n is given by

$$\begin{aligned}\langle n\rangle &= \mathrm{E}[n] = \sum_n p_m(n)n,\\ &= \sum_n\sum_N p_B(n|N,\varepsilon)p_P(N)n,\\ &= \sum_N p_P(N)\sum_n p_B(n|N,\varepsilon)n,\\ &= \varepsilon\sum_N p_P(N)N,\\ &= \varepsilon\langle N\rangle,\end{aligned} \tag{3.26}$$

where in the third line we commuted the sums, and in the fourth, we used the mean (Eq. 3.18) of the binomial distribution. The average measured multiplicity, $\langle n\rangle$, is thus, as expected, the product of the average produced multiplicity $\langle N\rangle$ by the efficiency ε. The second moment can be calculated in a similar fashion (see Problem 3.8):

$$\left\langle n^2\right\rangle = \varepsilon\left(1-\varepsilon\right)\langle N\rangle + \varepsilon^2\left\langle N^2\right\rangle. \tag{3.27}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

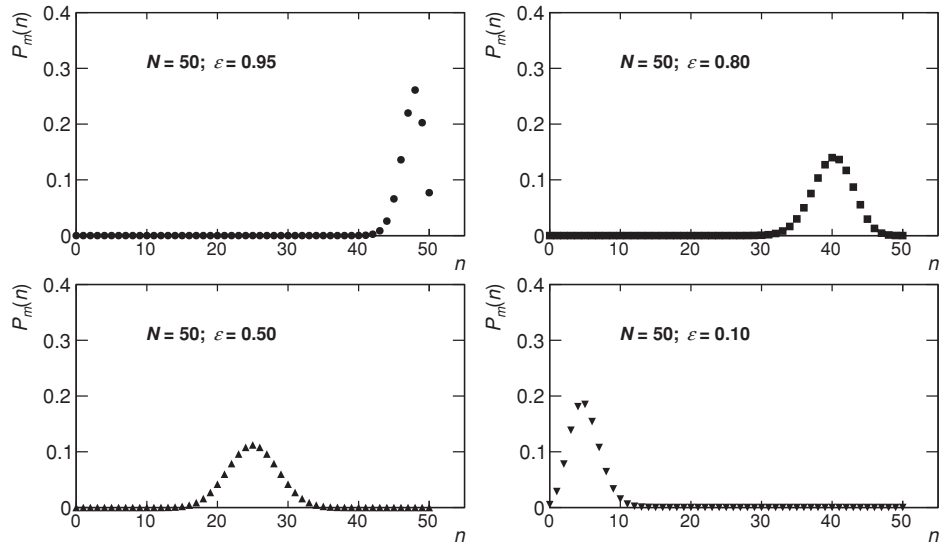


Fig. 3.3 Examples of multiplicity distributions obtained with efficiencies of 95%, 80%, 50%, and 10% for a produced multiplicity consisting of a single value of 50 particles.

The variance of the measured multiplicity is thus

$$\mathrm{Var}[n] = \langle n^2 \rangle - \langle n \rangle^2 \tag{3.28}$$

$$= \varepsilon^2 \mathrm{Var}[N] + \varepsilon\,(1-\varepsilon)\,\langle N \rangle\,. \tag{3.29}$$

While the average measured multiplicity is proportional to the produced multiplicity, the relationship between the variances of the measured and produced multiplicities is not as straightforward, given that it involves a term that is proportional to the multiplicity and a nontrivial dependence on the efficiency, $\varepsilon\,(1-\varepsilon)$. As we will discuss in further detail in Chapter 12, this implies that observables based on the variance of the multiplicity are not *robust* and cannot be trivially corrected for efficiency effects by taking simple ratios. However, as we shall see in that same chapter, factorial moments lend themselves naturally to such corrections.

3.1.5 Negative Binomial Distribution

The negative binomial distribution, as its name suggests, is intimately related to the binomial distribution. Here again, one considers successive trials that may yield successes S or failures $\overline{S}$. However, instead of considering the number of successes given n trials, one studies the number of trials completed before a prescribed number of success, k, is encountered. In this context, the last trial performed is by definition a success, and prior to this success, there must have been $k-1$ successes in the previous $n-1$ trials. The probability of k successes, with the last one being a success, thus yields the **negative binomial**

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

distribution, which may be written

$$p(n|k,\varepsilon) = \binom{n-1}{k-1} \varepsilon^k (1-\varepsilon)^{n-k}, \tag{3.30}$$

where ε denotes the probability of success S, as in the case of the regular binomial distribution.

The negative binomial distribution is particularly useful toward the description of produced particle multiplicity distributions measured in proton–proton collisions at high beam energy (see, e.g., [197] and references therein) or the number of photons produced by a blackbody at finite temperature [110].

3.1.6 A Technical Note

Computation of the binomial, multinomial (§3.2), and Poisson (§3.3) distributions requires the evaluation of the factorial of numbers that can be arbitrarily large. Although binomial and multinomial coefficients are finite, their evaluation can run into severe numerical instabilities, or floating point overflows, even with modern computers. Given that one is typically interested in ratios of factorial multiplied by powers of "p," it is often convenient to calculate the logarithms of the coefficients rather than their actual values. This can easily be accomplished using Stirling's[2] factorial formula:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n. \tag{3.31}$$

3.2 Multinomial Distribution

3.2.1 Definition

The **multinomial distribution** is an extension of the binomial distribution applicable when a measurement (random selection, or trial) can lead to many different outcomes of definite probability.

Consider, for instance, an "experiment" consisting of picking balls at random from a vat containing 25 red balls, 50 green balls, 100 yellow balls, and 25 black balls. The total number of balls is 200. Balls are replaced in the vat after they have been picked. The probability of picking one red ball in one trial equals the ratio of the number of red balls to the total number of balls in the vat, $p_{\text{red}} = 0.125$. Likewise, the probabilities to randomly pick a green, yellow, or black ball are $p_{\text{green}} = 0.25$, $p_{\text{yellow}} = 0.5$, and $p_{\text{black}} = 0.125$, respectively; and the sum of the four probabilities is unity. Next, consider the probability of picking two red balls and one green ball in one trial (consisting of a random selection of three balls). The probability of this outcome is obviously proportional to the product

[2] James Stirling, Scottish mathematician, 1692–1770.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

$p_{\text{red}}^2 p_{\text{green}}$, i.e., it is proportional to the product of probabilities of the different colors to a power equal to the number of balls of the given color being picked out.

Let us next consider, more generally, that a measurement could yield m types of objects or outcomes (e.g., four distinct colors, five types of particle species, etc.). Let us identify these outcomes with an index i spanning values $1, 2, \ldots, m$, and denote their respective probabilities p_i. By definition, the sum of these m probabilities equals unity:

$$\sum_{i=1}^{m} p_i = 1. \tag{3.32}$$

Next, consider the probability of getting $n_1, n_2, \ldots, n_m$ instances of type $i = 1, \ldots, m$ within a sample of N objects. The probability of a specific combination of $n_1, n_2, \ldots, n_m$ instances of each type is proportional to the product of their respective probabilities to the power n_i:

$$\text{Probability} \propto p_1^{n_1} \times p_2^{n_2} \times \cdots \times p_m^{n_m}. \tag{3.33}$$

There are many ways to generate a certain combination of $n_1, n_2, \ldots, n_m$ objects of types $i = 1, \ldots, m$. One must thus sum the probabilities of all these different sequences, while accounting for the fact that the order of the objects in a particular sequence is irrelevant. There are $N!$ possible sequences of N objects (outcomes) but since objects of a given type i are all equivalent (indistinguishable), one must divide the total number of sequences by the number of equivalent sequences of each type, $n_1! \ldots n_m!$. One then obtains the multinomial distribution

$$p_M(n_1, \ldots, n_m | N, p_1, \ldots, p_m) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} \times p_2^{n_2} \times \cdots \times p_m^{n_m}, \tag{3.34}$$

where by construction $\sum_{i=1}^{m} n_i = N$. It is often convenient to write the multinominal distribution using multinomial coefficients defined as follows:

$$\binom{N}{n_1, n_2, \ldots, n_m} \equiv \frac{N!}{n_1! n_2! \cdots n_m!}. \tag{3.35}$$

And using the product notation $\prod_{i=1}^{m} x_i = x_1 \times \cdots x_m$, one gets

$$p_M(n_1, \ldots, n_m | N, p_1, \ldots, p_m) = \binom{N}{n_1, \ldots, n_m} \prod_{i=1}^{m} p_i^{n_i}. \tag{3.36}$$

A derivation of the multinomial distribution may also be obtained with the Bayesian technique used in §3.1 for the derivation of the binomial distribution. In this case, the prior information I may be stated as follows: (1) there are m mutually exclusive and logically independent outcomes O_i, with $i = 1, \ldots, m$; (2) the outcome of a given draw is independent of the outcome of other draws. A draw, or event E, may thus be represented as a disjunction (i.e., an OR) of the m different outcomes, which in this context is equivalent to their sum:

$$E = O_1 + O_2 + \cdots + O_m. \tag{3.37}$$

A series of N draws, considered as an event M_N, may then be written

$$M_N = (O_1 + O_2 + \cdots + O_m)^N \tag{3.38}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.2 Properties of the multinomial distribution

Function	$p_M(n_1, \ldots, n_m \vert N; p_1, \ldots, p_m) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} \times p_2^{n_2} \times \cdots \times p_m^{n_m}$
Characteristic Fct	$\phi(t_1, t_2, \ldots, t_m) = \left(\sum_{j=1}^m p_j e^{it_j} \right)^N$
Moment Gen Fct	$M(t_1, t_2, \ldots, t_m) = \left(\sum_{j=1}^m p_j e^{t_j} \right)^N$
Moments	
1st	$\mu_i = \mu_i' = N p_i$
2nd	$\mu_{2,i}' = N p_i (1 - p_i + N p_i)$
Covariance	$\mathrm{Cov}[n_i, n_j] = -N p_i p_j$

Repeating the reasoning carried for the binomial distribution, we find that the probability of a specific draw depends on the number of occurrences of each type. An event may then be described as consisting of n_1 items of type O_1, n_2 items of type O_2, and so on, with $\sum_i n_i = n$. The number of ways such outcomes may occur is given by multinomial coefficients (3.35). The probability of a specific outcome with $(n_1, n_2, \ldots, n_m)$ is thus indeed given by Eq. (3.36).

3.2.2 Properties of the Multinomial Distribution

Moments of the multinomial distribution are easily calculated based on its characteristic function

$$\phi(t_1, t_2, \ldots, t_m) = \left(\sum_{j=1}^m p_j e^{it_j} \right)^N . \tag{3.39}$$

One readily verifies that the mean and variance of variables n_i have expressions similar to those found for the binomial distribution:

$$\mu_i = \mathrm{E}[n_i] = N p_i, \tag{3.40}$$

$$\sigma_i^2 = \mathrm{Var}[n_i] = N p_i (1 - p_i). \tag{3.41}$$

One must, however, also consider the correlation between different variables. This can be quantified by calculating the covariance, noted $\mathrm{Cov}[n_i, n_j]$, of two variables n_i and n_j such that $i \neq j$. The calculation of these covariances can also be carried out based on the characteristic function of the multinomial distribution (see Problem 3.7). One gets

$$\mathrm{Cov}[n_i, n_j] = -N p_i p_j. \tag{3.42}$$

The covariance is negative because the total number of objects sampled is fixed. If one value n_i increases, the others must decrease on average, and they therefore have a negative correlation.

The binomial and multinomial distributions are known as **exchangeable distributions** because they satisfy

$$p(O_j | O_{j-1}, I) = p(O_j | O_{j+1}, I), \tag{3.43}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

where O_j stands for the outcome of the jth draw. This equation may seem rather peculiar since it indicates that the probability of an outcome O_j, given a previous outcome O_{j-1}, is known to have occurred is equal to the probability of the same outcome given a subsequent outcome O_{j+1} is known to occur. How can this be causally possible? The answer is that it is not: the connection between these probabilities is purely logical, not causal. This peculiar property of Bayesian probabilities is discussed at length in §4.5 of ref. [97] . The exchangeable property is also shared by the **hypergeometric distribution**

$$p(r|N, M, n) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}}, \tag{3.44}$$

which expresses the probability of an event (a draw) involving r outcomes S in n tries from a pool of objects containing N objects, M of which are of type S and the remainder $N - M$ are of type $\overline{S}$ [117].

3.3 Poisson Distribution

3.3.1 Definition

The **Poisson distribution** is a discrete probability distribution that expresses the probability of a number of occurrences or events in a fixed interval of space or time. It is defined by the following prior information I or model: the probability of an event to occur in a given time interval $[t, t + dt]$ (or alternatively spatial interval $[x, x + dx]$) is defined as $p\ dt$, where p is a constant value independent of t (or x). Furthermore, the occurrence of an event at time t has no influence on the occurrence of events taking place in any other time intervals before or after the time t.

Let us define a predicate E as stating that no event occurred in the time interval $[0, t + dt]$ and let a function $q(t)$ represent the probability that no event occurred during the interval $[0, t]$. We then write

$$p(E|I) \equiv q(t + dt). \tag{3.45}$$

The statement E can be also formulated as the conjunction (i.e., an AND) between two propositions E_1 and E_2 such that E_1 states that no event occurred in the interval $[0, t]$, and E_2 states that no event occurred in the interval $[t, t + dt]$:

$$E = E_1, E_2 \tag{3.46}$$

We apply the product rule to express the probability of E in terms of the probabilities of E_1 and E_2:

$$p(E|I) = p(E_1, E_2|I) = p(E_1|I)p(E_2|E_1, I). \tag{3.47}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Since the two time intervals are mutually exclusive, and by virtue of the model hypothesis that the occurrence of an event at time t has no influence on the occurrence of events taking place in any other time intervals, we get

$$p(E|I) = p(E_1|I)p(E_2|I). \tag{3.48}$$

The probabilities $p(E|I)$ and $p(E_1|I)$ are quite obviously equal to $q(t+dt)$ and $q(t)$, respectively, while the probability $p(E_2|I)$ is equal to $1 - pdt$ by virtue of our first model hypothesis. The preceding expression thus becomes

$$q(t+dt) = q(t)\,(1 - pdt)\,. \tag{3.49}$$

Defining the difference $dq = q(t+dt) - q(t)$, we obtain the differential equation

$$\frac{dq}{dt} = -pq(t), \tag{3.50}$$

which holds solutions of the form

$$q(t) = q(0)\exp(-pt), \tag{3.51}$$

with $q(0) = 1$ as an obvious initial condition. This gives the exponential distribution (revisited in §3.5) expressing the probability that no events occurred in the time interval $[0, t]$ under the condition of the model (prior information I) defined previously.

Let us next consider a subsidiary proposition, noted C_n, that exactly n counts (or events) occurred at ordered times $t_1 < t_2 \ldots < t_n$ during the time interval $[0, t]$, and each within infinitesimal intervals $dt_1, dt_2, \ldots, dt_n$. The proposition may then be regarded as a conjunction of $2n+1$ simpler propositions, that there were no events in the interval $[0, t_1]$, one event during $[t_1, t_1 + dt_1]$, no events during $[t_1, t_2]$, one event during $[t_2, t_2 + dt_2]$, and so on. Applying once again the product rule and the model hypothesis that distinct time intervals are independent, one can write

$$\begin{aligned} p(C_n|p, I) &= \exp(-pt_1) \times pdt_1 \times \exp(-p(t_2 - t_1) \times pdt_2 \times \ldots \\ &\quad \times \exp(-p(t_n - t_{n-1}) \times pdt_n \times \exp(-p(t - t_n) \\ &= \exp(-pt)\, p^n dt_1 \ldots dt_n. \end{aligned} \tag{3.52}$$

In order to obtain the probability of observing exactly n events in the time interval $[0, t]$, irrespective of their exact times, we must sum the probabilities $p(C_n|p, I)$ by accounting for the fact that each of the n events may happen at any time in the interval $[0, t]$ while respecting the order $t_1 < t_2 \ldots < t_n$, that is, such that t_1 can occur anytime in $[0, t_2]$, t_2 can occur anytime in $[0, t_3]$, and so on.

$$p(n|p, t, I) = \exp(-pt)\, p^n \int_0^t dt_n \ldots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1. \tag{3.53}$$

The time ordered integral yields $t^n/n!$ and we thus obtain the Poisson distribution

$$p(n|p, t, I) = \frac{\exp(-pt)\,(pt)^n}{n!}. \tag{3.54}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

It is common practice to express the distribution in terms of the average number of counts in the finite time interval $[0, t]$, denoted $\lambda = pt$:

$$p_P(n|\lambda) = \frac{\exp(-\lambda)\,\lambda^n}{n!} \tag{3.55}$$

3.3.2 Properties of the Poisson Distribution

Let us first verify that the sum of the probabilities of all values of n yields unity:

$$\sum_{n=0}^{\infty} p_P(n|\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda}, \tag{3.56}$$

$$= e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}. \tag{3.57}$$

In the second line, the sum corresponds to a Taylor expansion of e^{λ}. Substituting this expression for the sum, we get the anticipated result:

$$\sum_{n=0}^{\infty} p_P(n|\lambda) = e^{-\lambda} e^{\lambda} = 1. \tag{3.58}$$

The distribution is thus properly normalized and qualifies for use as a PDF.

The moments of the Poisson distribution, $\mu'_k = \mathrm{E}[n^k]$, may be obtained by direct sum of Eq. (3.55), but are more conveniently calculated as derivatives of the characteristic function of the distribution which we proceed to evaluate:

$$\phi_n(t) = \mathrm{E}\left[e^{itn}\right] = \sum_{n=0}^{\infty} e^{itn} \frac{e^{-\lambda}\lambda^n}{n!}, \tag{3.59}$$

$$= e^{-\lambda} \sum_{n=0}^{\infty} \frac{\left(\lambda e^{it}\right)^n}{n!}. \tag{3.60}$$

The sum corresponds to the Taylor expansion of $e^{\lambda e^{it}}$. We thus obtain

$$\phi_n(t) = e^{\lambda(e^{it}-1)} \tag{3.61}$$

as the characteristic function of the Poisson distribution. The moments μ'_k of the Poisson distribution are then obtained from kth order derivatives of this expression

$$\mu'_k = i^{-k} \frac{d^k}{dt^k} \phi_n(t)\bigg|_{t=0}. \tag{3.62}$$

The first and second moments are

$$\mu'_1 = \mathrm{E}[n] = i^{-1} \frac{d}{dt} e^{\lambda(e^{it}-1)}\bigg|_{t=0} = \lambda, \tag{3.63}$$

$$\mu'_2 = \mathrm{E}[n^2] = i^{-2} \frac{d^2}{dt^2} e^{\lambda(e^{it}-1)}\bigg|_{t=0} = \lambda^2 + \lambda \tag{3.64}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.3 Properties of the Poisson distribution

Function	$p_P(n\|\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}$
Characteristic Fct	$\phi(t) = e^{\lambda\left(e^{it}-1\right)}$
Moment Gen Fct	$M(t) = e^{\lambda(e^t-1)}$
Moments	
1st	$\mu = \mu'_i = \lambda$
2nd	$\mu'_2 = \lambda\,(1+\lambda)$
3rd	$\mu'_3 = \lambda\left(1+3\lambda+\lambda^2\right)$
4th	$\mu'_4 = \lambda\left(1+7\lambda+6\lambda^2+\lambda^3\right)$
Central moments	
2nd	$\mu_2 = \lambda$
3rd	$\mu_3 = \lambda$
4th	$\mu_4 = \lambda\,(1+3\lambda)$
Skewness excess	$\gamma_1 = \lambda^{-1/2}$
Kurtosis excess	$\gamma_2 = \lambda^{-1}$

from which we obtained the variance (2nd centered moment) of the Poisson distribution:

$$\text{Var}[n] = \mu'_2 - (\mu'_1)^2 = \lambda. \tag{3.65}$$

The fact that the variance and the mean are equal is the hallmark of the Poisson distribution. It may be used as a rough diagnostic of measured distributions to establish whether they exhibit a **Poissonian behavior**. Higher moments, shown in Table 3.3, are easily obtained by calculating higher derivatives of Eq. (3.61).

The Poisson distribution is illustrated in Figure 3.4 for selected values of the parameter λ. Examples of application of the distribution are discussed in §§3.3.3 and 3.3.4. A derivation of the Poisson distribution as a limiting case of the binomial distribution is presented in §3.3.5 and its behavior in the large λ limit discussed in §3.3.6. The gamma distribution, presented in §3.6, may be regarded as an extension of the Poisson distribution for continuous values of n (i.e., $n \in \mathbb{R}$).

3.3.3 Example 1: Radioactive Decay Rates

The Poisson distribution is particularly useful in the context of processes (or systems) in which the probability of a certain phenomenon is small. For instance, consider repeated measurements of the number of decays of a radioactive source in a specific time interval Δt. The decay rate (activity) of the source is defined as the number of (nucleus) decays per time interval. It is proportional to the number N of nuclei present in the source and inversely proportional to the nucleus lifetime τ:

$$\frac{dN}{dt} = \frac{N}{\tau}. \tag{3.66}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

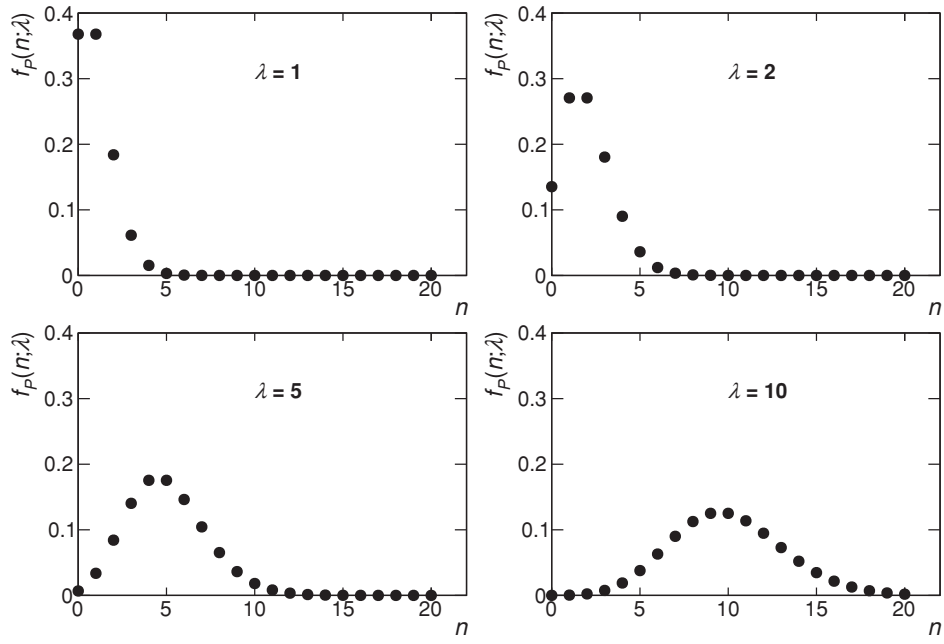


Fig. 3.4 Examples of the Poisson distribution with four different values of the parameter λ.

The expected (i.e., average) number of decays, $\langle n \rangle$, in a finite time interval Δt is thus

$$\langle n \rangle = \frac{dN}{dt}\Delta t = \frac{N}{\tau}\Delta t. \tag{3.67}$$

This result is strictly valid if N does not significantly vary during the time interval Δt and is thus applicable for sufficiently large N and small Δt. The decay of radioactive nuclei is, however, a stochastic process. While one can determine the average number of decays in a time interval of a specific duration, the number of decays in successive measurements of the number of decays in Δt will yield random numbers, n, that fluctuate about the mean $\langle n \rangle$. Given that decays are independent of one another, and for large sources (i.e., large N), one can model the probability of n decays during the time interval Δt with the Poisson distribution:

$$p_P(n; \langle n \rangle) = \frac{\langle n \rangle^n}{n!} e^{-\langle n \rangle}. \tag{3.68}$$

This implies it is possible to use the average of the number of decays, $\langle n \rangle$, observed over repeated measurements, to determine the lifetime τ of the nucleus even though it may be quite large relative to the duration of the experiment, or even the typical human lifespan. Uncorrelated fluctuations of the number n from measurement to measurement are commonly referred to as **Poisson fluctuations** or **Poisson noise**.

Although the Poisson distribution is defined for discrete values of the variable n, it can nonetheless be treated as a continuous function of a variable x, provided it is integrated over sufficiently large ranges $\Delta x \gg 1$. In fact, one finds that in the very large n limit, the

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Poisson distribution reduces to a Gaussian distribution (see §3.9 for a demonstration of this result).

3.3.4 Example 2: Detector Occupancy

Imagine you are instructed to design a time-of-flight hodoscope for a large particle physics experiment. You are told that one can expect an average of 100 particles in the acceptance of the hodoscope. You are then required to design the detector in such way that detector units of the hodoscope should have a hit occupancy, in other words, a rate of particle crossing each units, that does not exceed 5%. How many units or segments should the hodoscope then be made of?

Let us assume that, to first order, there exists no correlation between the particles entering the acceptance of the detector. The average occupancy of a detector segment is thus equal to the ratio, λ, of the average number of particles, $\langle n \rangle$, to the number of segments, N:

$$\lambda = \frac{\langle n \rangle}{N}. \tag{3.69}$$

One can use Poisson statistics to calculate the probability that a particular segment will be traversed by n particles in a given event:

$$P(n) = \frac{\lambda^n e^{-\lambda}}{n!}. \tag{3.70}$$

We wish to limit the probability of multiple occupancy of a segment to less than 5%.

$$P(n \geq 2) = 1 - P(0) - P(1) \leq 0.05. \tag{3.71}$$

Substituting the expression of the Poisson distribution, one obtains

$$\begin{aligned} P(\text{multiple occupancy}) &= 1 - e^{-\lambda} - \lambda e^{-\lambda}, \\ &\approx 1 - 1 + \lambda - \lambda + \lambda^2 + O(2), \\ &= \lambda^2 \leq 0.05. \end{aligned} \tag{3.72}$$

We then conclude the hodoscope should comprise at least $N = 447$ segments to limit the multiple hit occupancy to no more than 5%.

3.3.5 Poisson Distribution as a Limit of the Binomial Distribution

We derived the Poisson distribution as the probability of a given number of occurrences in a fixed interval of time, but it may also be obtained as a limit of the binomial distribution for $\varepsilon \to 0$, $N \to \infty$ with $\lambda = \varepsilon N$. This may be readily demonstrated based on the limiting behavior of the characteristic function of the binomial distribution. Recall from §2.10 that the characteristic function of a PDF can be considered as an alternative definition of the distribution. A limit of the binomial distribution may then be used to define a "new" probability distribution. We will quickly show this new distribution is in fact the Poisson distribution.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

One finds, by simple integration, that the characteristic function of the binomial distribution, Eq. (3.5), is given by

$$\phi(t) = \left[\varepsilon\left(e^{it} - 1\right) + 1\right]^N. \tag{3.73}$$

Let $z = \lambda\left(e^{it} - 1\right)$. Using the binomial theorem, Eq. (3.73) may then be written

$$\phi(t) = \left(1 + \frac{z}{N}\right)^N = \sum_{k=0}^{N} \frac{N!}{(N-k)!k!} \frac{z^k}{N^k}. \tag{3.74}$$

In the limit $N \to \infty$, this yields $\exp(z)$. One thus obtains

$$\phi(t) = \exp\left[\lambda\left(e^{it} - 1\right)\right], \tag{3.75}$$

which happens to be the characteristic function of the Poisson distribution. The Poisson distribution thus constitutes a limiting case of the binomial distribution for $\varepsilon \to 0, N \to \infty$ with $\lambda = \varepsilon N$.

$$\lim_{\substack{N\to\infty \\ \varepsilon\to 0}} p_B(n|N, p) = p_P(n|\lambda) = \frac{\lambda^n}{n!} e^{-\lambda}. \tag{3.76}$$

3.3.6 Large λ Limit of the Poisson Distribution

In the previous section, we saw that the Poisson distribution is a limiting case of the binomial distribution for $\varepsilon \to 0$, $N \to \infty$ and finite values of $\lambda = \varepsilon N$. It is also interesting to consider the behavior of the Poisson distribution $p_P(n|\lambda)$ in the limit $\lambda \to \infty$. Although the Poisson distribution is formally defined for discrete (integer) values of n, we can treat n as a continuous variable x in the limit of large λ as long as we integrate the function over an interval much larger than unity. Given the standard deviation of the Poisson is equal to the square root of λ, it is convenient and meaningful to define

$$x = \frac{n - \lambda}{\sqrt{\lambda}}, \tag{3.77}$$

which has, by construction, a mean of zero and a variance equal to unity. The characteristic function (3.75) of the PDF of x may then be written

$$\phi_x(t) = E\left[e^{itx}\right], \tag{3.78}$$

$$= E\left[e^{itn/\sqrt{\lambda}} e^{-it\sqrt{\lambda}}\right], \tag{3.79}$$

$$= e^{-it\sqrt{\lambda}} E\left[e^{itn/\sqrt{\lambda}}\right], \tag{3.80}$$

$$= e^{-it\sqrt{\lambda}} \quad \phi_n\left(\frac{t}{\sqrt{\lambda}}\right), \tag{3.81}$$

where by definition ϕ_n is the characteristic function of the Poisson distribution. We substitute the expression of the characteristic function of the Poisson distribution in the preceding and get

$$\phi_x(t) = \exp\left[\lambda\left(e^{it/\sqrt{\lambda}} - 1\right) - it\sqrt{\lambda}\right]. \tag{3.82}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

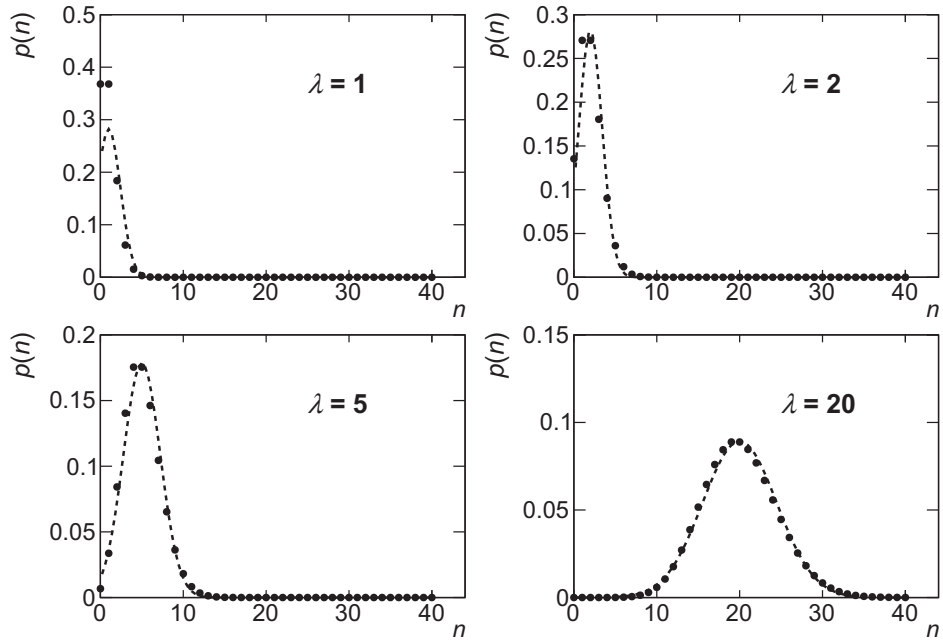


Fig. 3.5 Illustration of the convergence of the Poisson distribution (solid circle) with the normal distribution (dashed line) in the "large" λ limit. Normal distributions are plotted for $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$.

Expansion of the innermost exponential yields $1 + \frac{it}{\sqrt{\lambda}} + \frac{1}{2}(\frac{it}{\sqrt{\lambda}})^2 + O(3)$ in the limit $\lambda \to \infty$. Equation (3.82) thus reduces to

$$\phi_x(t) = \exp\left(-\tfrac{1}{2}t^2\right), \tag{3.83}$$

which is the expression of the characteristic function of the standard normal distribution already introduced in §2.10.2 and further discussed in §3.9. One thus concludes that for large values of λ, the variable n follows a Gaussian distribution. In other words, for large values of λ, the Poisson distribution may be approximated by a Gaussian distribution with mean and variance both equal to λ. Figure 3.5 illustrates how the Poisson distribution quickly converges toward the Gaussian distribution with mean $\mu = \lambda$ and standard deviation $\sigma = \sqrt{\lambda}$ for increasing values of the parameter λ.

3.4 Uniform Distribution

3.4.1 Definition

The **uniform distribution**, as its name suggests, implies all values of a continuous variable, x, are equally probable in a given range $[\alpha, \beta]$. It is defined in the range $[-\infty, \infty]$ as

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

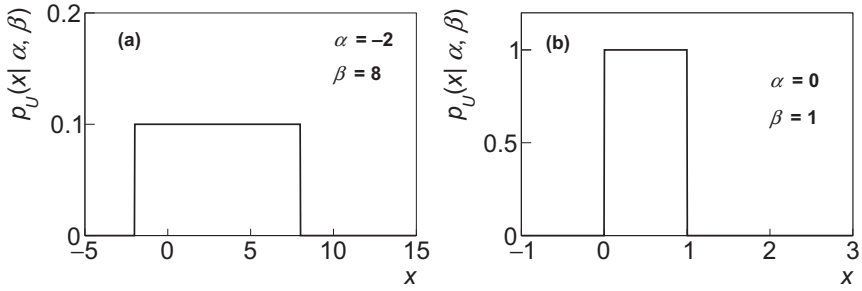


Fig. 3.6 (a) Example of the uniform distribution, $p_u(x|\alpha, \beta)$ with $\alpha = -2.0, \beta = 8.0$. (b) Standard uniform distribution.

follows:

$$p_u(x|\alpha, \beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \le x \le \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{3.84}$$

Figure 3.6 displays an example of the uniform distribution with values $\alpha = -2$ and $\beta = 8$. The probability density has a constant value of 0.1 in the range $-2 \le x \le 8$, and is null otherwise. The area under the curve equals unity and corresponds to the probability of observing x in the range $-2 \le x \le 8$: all values of x in that range are equally probable, and those outside of the range are *impossible*.

The cumulative density function (CDF) of p_u is given by

$$P_u(x) = \int_{-\infty}^{x} p_u(x'|\alpha, \beta)\, dx', \tag{3.85}$$

$$= \begin{cases} 0 & \text{for } x < \alpha, \\ (x-\alpha)/(\beta-\alpha) & \text{for } \alpha \le x \le \beta, \\ 1 & \text{for } x > \beta. \end{cases} \tag{3.86}$$

Setting $\alpha = 0$, and $\beta = 1$, one gets the **Standard Uniform Distribution**, $p_u(x|0, 1)$.

3.4.2 Properties of the Uniform Distribution

The mean, μ, of the uniform distribution is obtained by calculating the expectation value of x for this PDF.

$$\mu \equiv \mathrm{E}[x] = \int_{-\infty}^{\infty} p_u(x|\alpha, \beta) x\, dx, \tag{3.87}$$

$$= \frac{1}{\beta-\alpha} \int_{\alpha}^{\beta} x\, dx = (\beta+\alpha)/2. \tag{3.88}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.4 Properties of the uniform distribution

Function	$p_u(x\|\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$
Characteristic Fct	$\phi(t) = E\left[e^{itx}\right] = \left(e^{it\beta} - e^{it\alpha}\right)/(it\,(\beta-\alpha))$
Moment Gen Fct	$M(t) = E\left[e^{tx}\right] = \frac{e^{t\beta}-e^{t\alpha}}{t(\beta-\alpha)}$
Mode	Anywhere in the range $\alpha \le x \le \beta$
Median	$(\alpha+\beta)/2$
Moments	
1st	$\mu = \mu_i' = \frac{\beta+\alpha}{2}$
2nd	$\mu_2' = \frac{\alpha^2+\alpha\beta+\beta^2}{3}$
kth	$\mu_k' = \frac{1}{k+1}\sum_{i=0}^{k}\alpha^i\beta^{k-i}$
Central moments	
2nd	$\sigma^2 \equiv \mu_2 = \frac{(\beta-\alpha)^2}{12}$
Skewness excess	$\gamma_1 = 0$
Kurtosis excess	$\gamma_2 = 6/5$

The variance, σ^2, is readily obtained by calculating $\mathrm{E}[x^2] - \mathrm{E}[x]^2$ as follows (see Problem 3.13):

$$\sigma^2 \equiv \mathrm{Var}[x] = \frac{1}{\beta-\alpha}\int_{-\infty}^{\infty}(x-\mu)^2 p_u(x|\alpha,\beta)\,dx, \tag{3.89}$$

$$= \frac{(\beta-\alpha)^2}{12}. \tag{3.90}$$

From this expression, we conclude that a flat distribution of width L has a standard deviation $L/\sqrt{12}$. The skewness, the kurtosis excess, and other higher moments may be obtained similarly and are listed in Table 3.4.

3.4.3 Random Number Generation

The uniform and standard uniform distributions find use in a variety of applications, and most particularly, random number generation.

Given a set of random numbers $\{r_i\}$, with $i = 1, \ldots, n$ distributed according to the standard uniform distribution $p_u(x|0,1)$, one can trivially obtain random numbers following a uniform (but nonstandard) distribution $p_u(x|\alpha,\beta)$ using a simple linear transformation as follows:

$$r_i' = \alpha + (\beta-\alpha)\,r_i. \tag{3.91}$$

In fact, pseudo-random number generators (defined in Chapter 11) provided with popular mathematical software libraries (e.g., Mathematica®, MATLAB®, ROOT) typically feature a "core" random number generator that produces a standard uniform distribution, which may then be transformed, according to Eq. (3.91), to obtain arbitrary uniform distributions.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

The standard uniform distribution can also be transformed to generate random numbers according to a wide variety of less trivial PDFs. For instance, consider the function $f(x)$ and its cumulative distribution, $y = F(x)$:

$$y = F(x) = \int_{-\infty}^{x} f(x')\,dx'. \tag{3.92}$$

By construction, one has

$$\frac{dy}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f(x')\,dx' = f(x). \tag{3.93}$$

As we saw in §2.6, the probability of an event is invariant under a change of variable $x \to y$. Let $g(y)$ represent the PDF of y, that is, the probability of observing y in the range $[y, y + dy]$. One can then use Eq. (2.60) to express $g(y)$ in terms of $f(x)$. One gets

$$g(y) = f(x)\left|\frac{dx}{dy}\right| = f(x)\left|\frac{dy}{dx}\right|^{-1} = 1, \tag{3.94}$$

and we conclude that $g(y)$ is a uniform PDF. This implies that any PDF $f(x)$ of a continuous variable may in principle be expressed (and calculated) in terms of a uniform distribution. This property is particularly important for the generation of random numbers discussed in §13.3.3.

3.4.4 Example: Position Resolution of a Basic Particle Hit Detector

Magnetic spectrometers are used to measure the momentum of charged particles produced by nuclear collisions. In the bending plane, charged particles follow circular trajectories. The radius of these trajectories, or tracks, is proportional to the momentum of the particles and inversely related to the magnetic field (B). A precise knowledge of the B field therefore enables a measurement of the particle momenta. This requires a determination of the radius of curvature of the tracks, which is often accomplished by measuring "hit" positions along the track, in other words, positions where the particles traverse specific detector planes. In the most basic spectrometers, these detector planes are equipped with a collection of contiguous segments that can each detect the passage of a single particle. As a particle passes through a given segment or unit, it deposits energy and generates a signal. Assuming a particle can generate a signal or "hit" only in one segment, one finds that the hit position resolution is intrinsically limited by the width of the unit. The hit position is typically assumed to be the center of the segment. That means there is an uncertainty on the position proportional to the width of the segment. Let us estimate this error on the position by assuming that the particle has an a priori equal probability of crossing the unit anywhere along its width:

$$P(x) = \frac{1}{\Delta L}, \tag{3.95}$$

where ΔL is the actual width of the unit. By construction, the average position, $\langle x \rangle$, shall be the center of the unit. We can then use the standard deviation of the uniform distribution

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

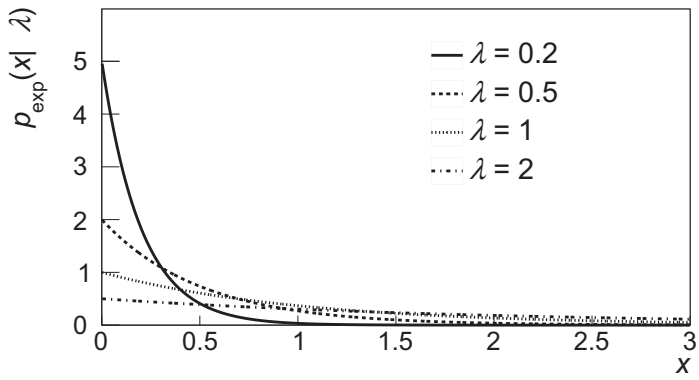


Fig. 3.7 Exponential distribution $p_{\text{exp}}(x|\lambda)$ for selected values of λ.

as an estimate of the error, σ_x, on the position:

$$\sigma_x = \text{Var}[x] = \int_0^{\Delta L} (x - \langle x \rangle)^2 \, P(x) \, dx = \frac{\Delta L}{\sqrt{12}}. \tag{3.96}$$

This simple estimate of the error can be used in a wide variety of contexts to estimate errors on measurements where the probability density is (approximately) uniform.

3.5 Exponential Distribution

3.5.1 Definition

The **exponential probability distribution** is defined as follows:

$$p_{\text{exp}}(x|\lambda) = \begin{cases} \lambda^{-1} \exp(-x/\lambda) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases} \tag{3.97}$$

It is determined by a single parameter, λ, which sets the falloff rate of the exponential, as illustrated in Figure 3.7.

The factor λ^{-1} ensures the proper normalization of $p_{\text{exp}}(x|\lambda)$ as a PDF. Indeed, calculating the cumulative distribution function of $p_{\text{exp}}(x|\lambda)$, one finds

$$P_{\text{exp}}(x|\lambda) = \int_{-\infty}^{x} p_{\text{exp}}(x'|\lambda) \, dx', \tag{3.98}$$

$$= \begin{cases} 1 - \exp(-x/\lambda) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0, \end{cases} \tag{3.99}$$

which converges to unity in the limit $x \to \infty$.

3.5.2 Interesting Features of the Exponential Distribution

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

As we saw in §3.3, the exponential distribution naturally arises in the context of stochastic phenomena where the probability of an event to occur in a given time interval $[t, t + dt]$ has a fixed value pdt where p is a constant value independent of t, and such that the occurrence of an event at time t has no influence on the occurrence of events taking place in any other time intervals before or after the time t. The event rate may then be described in terms of a constant τ^{-1} that specifies the average rate of events per unit of time. The Poisson distribution then corresponds to the probability of observing N events in a given time interval Δt:

$$p(N|\Delta t) = \frac{1}{N!}\,(t/\tau)^N\, e^{-t/\tau}, \tag{3.100}$$

and the probability of having no events, that is, $N = 0$, during an interval Δt is given by the exponential distribution $e^{-t/\tau}$.

The exponential PDF is useful in a variety of contexts. It can, for instance, describe the probability of observing the decay of a specific nuclear isotope as a function of time, t. The quantity λ then corresponds to the lifetime of the isotope and is usually noted τ instead. The exponential PDF may also be used to represent the probability of weakly decaying particles (such as D mesons) to decay at a certain distance from their point of production.

It is interesting to consider the time interval T between two events relative to the time interval Δt. Since no events occur between the two events for $T > \Delta t$, we can write

$$p(T > \Delta t) = e^{-\Delta t/\tau}. \tag{3.101}$$

Additionally, given the expression of the CDF, Eq. (3.98), one can also conclude that the probability of observing two events separated by a time T smaller than Δt is given by

$$p(T \leq \Delta t) = 1 - e^{-\Delta t/\tau}, \tag{3.102}$$

since the right-hand side of Eq. (3.102) is the complement of the probability $p(T > \Delta t)$.

The exponential distribution is often said to have "no memory." For instance, if no event occurred until a time t_1, the probability of having no event in a subsequent time t_2 is independent of t_1. For a fixed value t_1, this may be written

$$p(T > t_1 + t_2 | T > t_1) = \frac{e^{-(t_1+t_2)/\tau}}{e^{-t_1/\tau}} = e^{-t_2/\tau} = p(T > t_2), \tag{3.103}$$

which indeed implies that the probability of $T > t_1 + t_2$ does not depend on t_1. Since the probability of an event taking place in a certain time interval does not depend on what happened during prior time intervals, the exponential can hence be said to have no memory. This should come as no surprise, however, as per the derivation of the distribution presented in §3.3.1 in terms of independent events.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.5 Properties of the exponential distribution

Function	$p_{\exp}(x\|\lambda) = \lambda^{-1}\exp(-x/\lambda)$
Characteristic Fct	$\phi(t) = (1 - i\cdot\lambda t)^{-1}$
Moment Gen Fct	$M(t) = (1-\lambda t)^{-1}$
Mode	0
Median	$\ln(2)\lambda$
Moments	
1st	$\mu = \mu' = \lambda$
2nd	$\mu_2' = 2\lambda^2$
nth	$\mu_n' = \lambda^n n!$
Central moments	
2nd	$\sigma^2 \equiv \mu_2 = \lambda^2$
nth	$\mu_n = \Gamma(n+1,-1)\lambda^n/e$
Skewness excess	$\gamma_1 = 2$
Kurtosis excess	$\gamma_2 = 6$

3.5.3 Moments of the Exponential Distribution

The expectation value of x is (see Problem 3.14)

$$\mu = \mathrm{E}[x] = \frac{1}{\lambda}\int_0^\infty x\mathrm{e}^{-x/\lambda}\,dx = \lambda, \tag{3.104}$$

and the variance is given by

$$\mathrm{Var}[x] = \frac{1}{\lambda}\int_0^\infty (x-\lambda)^2\mathrm{e}^{-x/\lambda}\,dx = \lambda^2. \tag{3.105}$$

Other moments of the distribution are listed in Table 3.5. More generally, one can show (see Problem 3.14) that the moments of the exponential distribution are given by

$$\mu_n' = \lambda^n n! \tag{3.106}$$

whereas the central moments are given by

$$\mu_n = \frac{\Gamma(n+1,-1)}{e\lambda^n}, \tag{3.107}$$

where $\Gamma(a,b)$ is an incomplete gamma function.

3.5.4 Example: Radiocarbon Decay and Half-Life

A convenient and useful way to visualize the exponential PDF is by means of the concept of half-life. Consider carbon-14 (^{14}C), an unstable carbon isotope consisting of six protons and eight neutrons. Commonly referred to as radiocarbon, ^{14}C decays spontaneously into a nucleus of nitrogen-14 (^{14}N, which consists of seven protons and seven neutrons) by

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

emission of an electron and an anti-neutrino. This decay is governed by the weak nuclear force. It involves the transmutation of one of the eight neutrons of the ^{14}C nucleus into a proton, the production of an electron (which conserves the total electric charge), and the emission of a neutrino (which conserves energy and momentum):

$$^{14}_{6}\mathrm{C} \rightarrow ^{14}_{7}\mathrm{N} + e^{-} + \overline{\nu}_e \tag{3.108}$$

The isotope ^{14}C is naturally present in all organic materials. It forms the basis of the radiocarbon dating method used to determine the age of archaeological, geological, and hydrogeological artifacts. The isotope ^{14}C has a half-life ($t_{1/2}$) of 5,730 years. This means that given a sample of n ^{14}C nuclei, it should take (on average) 5,730 years for half of these nuclei ($n/2$) to be decayed. After an additional 5,730 years, one-half of the remaining half would have decayed, thereby leaving us with one-fourth ($n/4$) of the original amount. After three half-lives, one-eighth remains; after four half-lives, one-sixteenth is left, and so on. This can be written

$$n_{\text{undecayed}}(t) \sim \left(\frac{1}{2}\right)^{t/t_{1/2}}, \tag{3.109}$$

or equivalently

$$n_{\text{undecayed}}(t) \sim \exp\left(\frac{-t}{t_{1/2}/\ln 2}\right). \tag{3.110}$$

The quantity $t_{1/2}/\ln 2$ is readily identified as the mean life, τ, of the isotope:

$$\tau = t_{1/2}/\ln 2 = t_{1/2}/0.693147. \tag{3.111}$$

The half-life of ^{14}C, 5,730 years, makes it ideal for "dating" archeological samples that are a few to several thousand of years old. In practice, the technique appears to be limited to about 10 half-lives, corresponding to roughly 60,000 years.

3.6 Gamma Distribution

3.6.1 Definition

The **gamma distribution** (not to be confused with the gamma function, $\Gamma(x)$, see later) is encountered in a wide variety of situations and fields including econometrics, life sciences, as well as physical sciences. As such, it is commonly represented in terms of (at least) three distinct parameterizations in the scientific literature. In this book, we will use the notation most popular in Bayesian statistics, which involves a shape parameter, α, and a rate parameter β:

$$p_\gamma(x|\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)} \quad \text{for } x \geq 0 \;\; \alpha,\beta > 0, \tag{3.112}$$

where $\Gamma(x) = \int_0^\infty z^{n-1}\exp(-z)dz$, is the **gamma function** for $x > 0$. One readily verifies that the exponential distribution (§3.5) and the χ^2-distribution (§3.9) are two special cases

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

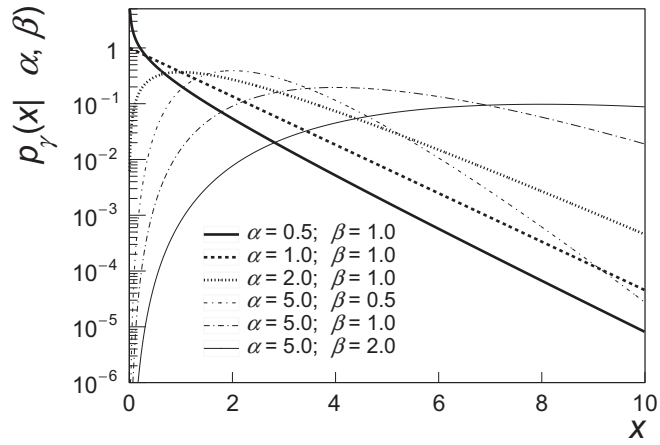


Fig. 3.8 Gamma distributions $p_\gamma(x|\alpha,\ \beta)$ for selected values of α and β.

of the gamma distribution, while integer values for the shape parameter α yield **Erlang distributions**, which corresponds to sums of α exponentially distributed random deviates, each with a mean $1/\beta$.

As illustrated in Figure 3.8, the gamma distribution spans a wide range of shapes and forms determined by the parameters α and β. It is thus quite convenient for modeling the PDF of random variables and is used in a wide variety of applications and systems.

Integration of p_γ in the interval $[0, x]$ yields the CDF denoted:

$$P_\gamma(x|\alpha, \beta) = \int_0^x p_\gamma(z|\alpha, \beta)\, dz = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} \tag{3.113}$$

where $\gamma(\alpha, \beta x)$ is known as **incomplete gamma function**. The CDF function $P_\gamma(x|\alpha, \beta)$ is commonly called **regularized gamma function**. It can be expressed in closed form if α is a positive integer. Indeed, since p_γ vanishes, by construction, both at $x = 0$ and for $x \to \infty$, integration by part yields:

$$P_\gamma(x|\alpha, \beta) = 1 - \sum_{k=0}^{\alpha-1} \frac{(\beta x)^k}{k!} \exp(-\beta x) \tag{3.114}$$

$$= \exp(-\beta x) \sum_{k=\alpha}^{\infty} \frac{(\beta x)^k}{k!}, \tag{3.115}$$

where in the second line, we substituted the complement of the partial Taylor series of $\exp(\beta x)$ found in the first line.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.6 Properties of the gamma distribution

Function	$p_\gamma(x\|\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)}$
Characteristic Fct	$\phi(t) = (1 - it/\beta)^{-\alpha}$
Moment Gen Fct	$M(t) = (1 - t/\beta)^{-\alpha}\ t < \beta$
Mode	$(\alpha - 1)/\beta \quad \text{for} \quad \alpha \geq 1$
1st Moment	$\mu = \mu' = \alpha/\beta$
2nd Central moment	$\sigma^2 \equiv \mu_2 = \alpha/\beta^2$
Skewness excess	$\gamma_1 = 2/\sqrt{\alpha}$
Kurtosis excess	$\gamma_2 = 6/\alpha$

3.6.2 Moments of the Gamma Distribution

Moments and other basic properties of the gamma distribution are listed in Table 3.6.

3.7 Beta Distribution

3.7.1 Definition

The **beta distribution** is defined as

$$p_\beta(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}\,(1-x)^{\beta-1}\,, \tag{3.116}$$

over the domain [0, 1] for real exponent values $\alpha, \beta > 0$ and as such may be regarded as an extension of the binomial distribution.

3.7.2 Properties of the Beta Distribution

Figure 3.9 illustrates how judicious choices of the **shape parameters** α and β produce functions that span a rather wide range of functional shapes in the domain $0 < x < 1$. As such, the beta distribution is extremely convenient and often used toward the expression of prior probabilities in Bayesian statistics (§7.2.2).

Remarkably, one finds that if a likelihood function is a binomial distribution, multiplying the prior by the likelihood yields another beta distribution. The beta distribution is then said to be a **conjugate prior** of the binomial distribution. In general, in cases in which the prior and posterior are members of the same family of distributions, with a given type of likelihood function, ones considers the prior and the likelihood to be **conjugate distributions**. Other cases of conjugate distributions include the normal distribution when used both as prior and likelihood functions, as well as the gamma distribution when used a prior for a Poisson likelihood. A more in-depth discussion of conjugate priors is presented in §7.3.3.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

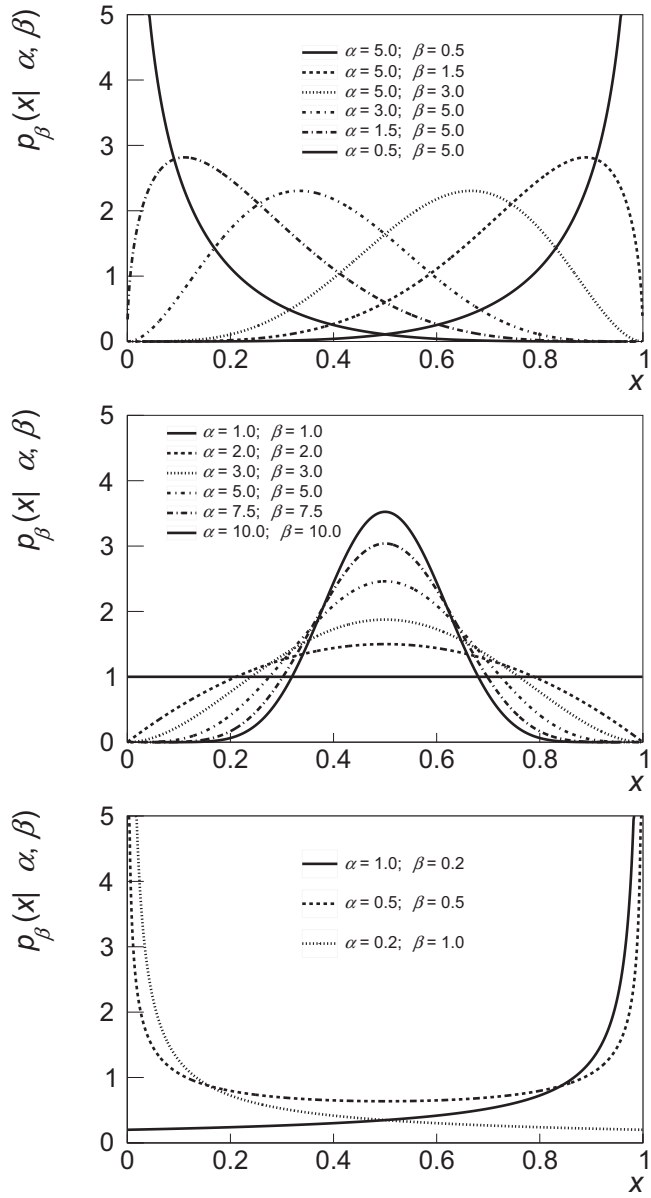


Fig. 3.9 Examples of the beta distribution $p_\beta(x|\alpha,\ \beta)$ for selected values of α and β.

Basic properties of the beta distribution are listed in Table 3.7. The moment-generating function of the beta distribution is calculated in terms of the expectation value of e^{tx}

$$M_X(\alpha, \beta, t) = \mathrm{E}[e^{tx}] = \int_0^1 e^{tx} p_\beta(x|\alpha, \beta)\, dx. \tag{3.117}$$

This expression corresponds to the confluent hypergeometric function

$$M_X(\alpha, \beta, t) = {}_1F_1(\alpha, \alpha + \beta, t), \tag{3.118}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.7 Properties of the beta distribution

Function	$p_\beta(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Characteristic Fct	${}_1F_1(\alpha, \alpha+\beta, it)$
Moment Gen Fct	${}_1F_1(\alpha, \alpha+\beta, t)$
Mode	$(\alpha-1)/(\alpha+\beta-2)$ for $\alpha, \beta > 1$
Median	$\approx (\alpha-1/3)/(\alpha+\beta-2/3)$ for $\alpha, \beta > 1$
1st Moment	$\mu = \alpha/(\alpha+\beta)$
2nd Central moment	$\text{Var}[x] = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$
Skewness	$\gamma_1 = 2/(\beta-\alpha)\sqrt{\alpha+\beta+1}/(\alpha+\beta+2)\sqrt{\alpha\beta}$
Kurtosis excess	$\gamma_2 = 6\left[(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)\right] / \alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)$

which may be written as an infinite series [5]

$$M_X(\alpha, \beta, t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{j=0}^{k-1} \frac{\alpha+j}{\alpha+\beta+j} \right) \frac{t^k}{k!}. \tag{3.119}$$

Similarly, the characteristic function of the beta distribution may be shown to be equal to

$$\phi_X(\alpha, \beta, t) = {}_1F_1(\alpha, \alpha+\beta, it) \tag{3.120}$$

$$= 1 + \sum_{k=1}^{\infty} \left(\prod_{j=0}^{k-1} \frac{\alpha+j}{\alpha+\beta+j} \right) \frac{(it)^k}{k!}. \tag{3.121}$$

3.8 Dirichlet Distributions

3.8.1 Definition

Much like the beta distribution may be regarded as an extension of the binomial distribution, the **Dirichlet distribution** may be viewed as an extension of the multinomial distribution. It is defined as a function of k continuous variables x_i and α_i according to

$$p_D(\vec{x}|\vec{\alpha}) = \frac{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \cdots x_k^{\alpha_k-1}, \tag{3.122}$$

with the following constraints:

$$\vec{x} = (x_1, x_2, \ldots, x_k) \qquad 0 < x_i < 1; \qquad \sum_{i=1}^{k} x_i = 1,$$

$$\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k) \qquad \alpha_i > 0; \qquad \sum_{i=1}^{k} \alpha_i \equiv \alpha,$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.8 Properties of the Dirichlet distribution

Function	$f(\vec{x}\|\vec{\alpha}) = \frac{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \cdots x_k^{\alpha_k-1}$
Mode	$x_i^{Mode} = (\alpha_i - 1)/(\sum_{j=1}^{k} \alpha_j - k)$ for $\alpha_i > 1$
Mean	$E[x_i] = \alpha_i / \sum_{j=1}^{k} \alpha_j$
Variance	$\mathrm{Var}[x_i] = \alpha_i (\alpha - \alpha_i) / \alpha^2 (\alpha + 1)$
Covariance	$\mathrm{Cov}[x_i, x_j] = -\alpha_i \alpha_j / \alpha^2 (\alpha + 1)$ for $i \neq j$
Note:	$\alpha = \sum_{i=1}^{k} \alpha_i$

where $\Gamma(z)$ is the regular gamma function. The Dirichlet distribution simplifies to a beta distribution for $k = 2$. Also note that the constraint $\sum_{i=1}^{k} x_i = 1$ implies the distribution is effectively $(k - 1)$ dimensional.

The Dirichlet distribution find applications, much like the beta distribution, in the context of Bayesian inference with finite statistics (see, e.g., §12.3.8).

Marginalization of the Dirichlet distribution, that is, reduction to one variable x_i, yields the beta distribution with parameters $r = \alpha_i$ and $s = \alpha - \alpha_i$.

$$p_\beta(x_i | r = \alpha_i, s = \alpha - \alpha_i) = \frac{x_i^{\alpha_i - 1} (1 - x_i)^{\alpha - \alpha_i - 1}}{\beta(\alpha_i, \alpha - \alpha_i)}. \tag{3.123}$$

3.8.2 Properties of the Dirichlet Distribution

The properties of the Dirichlet distribution are summarized in Table 3.8.

3.9 Gaussian Distribution

3.9.1 Definition

The **Gaussian distribution** is defined in terms of a continuous variable x in the domain $[-\infty, \infty]$:

$$p_G(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{3.124}$$

The distribution is given many names: statisticians and mathematicians use the term **normal distribution** while social scientists commonly refer to it as the **bell curve** because of its curved flaring shape. Physicists typically prefer the appellation **Gaussian distribution** in honor of Carl Friedrich Gauss[3] (1777–1855). Indeed, although the distribution was likely "discovered" by De Moivre as an approximation of the binomial distribution, and

[3] German mathematician and scientist who made major contributions in number theory, statistics, analysis, differential geometry, geodesy, geophysics, and physics. Gauss is considered by many as the greatest mathematician since antiquity.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

subsequently used by Laplace in 1783 to study measurement errors, its first extensive application in a scientific context is commonly attributed to Gauss, who, in 1809, used the distribution toward the analysis of astronomical data.

The name and qualifier **normal** is perhaps best suited for the distribution. Indeed, one finds that many distributions tend to the normal distribution in the large N limit. One also finds that the sum of a very large number of random variables has a normal distribution. This fact is embodied in the central limit theorem, which we discussed in §2.10.4. The normal distribution thus, indeed, describes that which is common or "normal." But given this text is targeted primarily toward physical scientists, we will hereafter mostly use the Gaussian distribution appellation commonly in vogue in physics.

Note that statisticians often use the notation $X \sim \mathcal{N}(\mu, \sigma^2)$ to indicate that a variable X is distributed according to a normal distribution of mean μ and variance σ^2. The distribution may also be defined in terms of a **precision parameter** ξ defined as the multiplicative inverse of the variance

$$\xi = \frac{1}{\sigma^2}, \tag{3.125}$$

and is then written

$$p_{\mathrm{G}}(x|\mu, \xi) = \sqrt{\frac{\xi}{2\pi}} \exp\left[-\frac{1}{2}\xi\,(x-\mu)^2\right]. \tag{3.126}$$

3.9.2 Moments of the Gaussian Distribution

We first note that while the function $p_{\mathrm{G}}(x|\mu, \sigma^2)$ does not have an analytical primitive (indefinite integral), its definite integral over the domain $[-\infty, \infty]$ nonetheless exists and can be shown to equal unity. The function is thus properly defined and usable as a probability density. This stems from the inclusion of the normalization factor $\sqrt{2\pi}\sigma$ (see Problem 3.15):

$$\int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = \sqrt{2\pi}\sigma. \tag{3.127}$$

One verifies, through integration by parts, that the two parameters μ and σ correspond to the mean and the variance of the distribution, respectively:

$$\mathrm{E}[x] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] x\,dx = \mu, \tag{3.128}$$

and

$$\mathrm{Var}[x] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] (x-\mu)^2\,dx = \sigma^2. \tag{3.129}$$

It is immaterial whether one considers σ, σ^2, or ξ, as the parameter determining the width of the distribution. However, given the Gaussian distribution is commonly used to

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

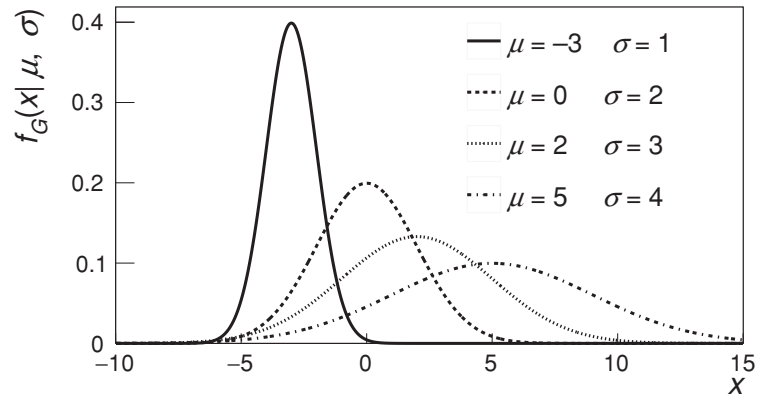


Fig. 3.10 Gaussian distribution $p_G(x|\mu, \sigma^2)$ for selected values of μ and σ^2.

describe fluctuations in the outcome of measurements, the precision label given to ξ is both practical and advantageous: indeed, as ξ increases, the width of the distribution decreases and thus reflects a measurement of greater precision. Evidently, σ^2 and σ are also both useful given they respectively represent the variance and standard deviation of the distribution.

Figure 3.10 displays examples of the Gaussian distribution for various combinations of μ and σ. Higher moments of the distribution are best calculated using the characteristic function (or moment-generating function) of the distribution listed in Table 3.9 (see Problem 3.16).

A special case of the Gaussian distribution, known as **standard normal distribution**, is obtained by setting $\mu = 0$ and $\sigma = 1$:

$$\varphi(x) \equiv p_G(x|\mu = 0, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{3.130}$$

The cumulative distribution of $\varphi(x)$, noted $\Phi(x)$, is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-x'^2/2\right) dx' = \frac{1}{2}\left[1 + \operatorname{erf}\left(x/\sqrt{2}\right)\right], \tag{3.131}$$

where $\operatorname{erf}(x)$ is the **error function**, defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{3.132}$$

It is also useful to be familiar with the **complementary error function** $\operatorname{erfc}(x)$, defined as

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt. \tag{3.133}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.9 Properties of the Gaussian (normal) distribution

Function	$p_G(x;\mu,\sigma^2) = (2\pi)^{-1/2}\sigma^{-1}\exp\left(-(x-\mu)^2/2\sigma^2\right)$
Characteristic Fct	$\phi(t) = \exp\left(i\mu t - \sigma^2 t^2/2\right)$
Moment Gen Fct	$M(t) = \exp\left(\mu t + \sigma^2 t^2/2\right)$
Mode	μ
Median	μ
Moments	
1st (Mean)	$\mu = \mu'$
2nd	$\mu'_2 = \mu^2 + \sigma^2$
3rd	$\mu'_3 = \mu^3 + 3\mu\sigma^2$
4th	$\mu'_4 = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$
Centered moments	
2nd (Variance)	σ^2
3rd	0
4th	σ^4
Skewness excess	$\gamma_1 = 0$
Kurtosis excess	$\gamma_2 = 1$

Additionally, note that the error function is odd in x:

$$\text{erf}(-x) = -\text{erf}(x) \tag{3.134}$$

The error function, erf(x), and the complementary error function, erfc(x), cannot be calculated analytically but are tabulated in many reference books. They are also provided in most mathematical software packages (e.g., Mathematica®, MATLAB®, and ROOT).

An arbitrary Gaussian distribution, $p_G(y;\mu,\sigma^2)$, function off a variable y with a mean μ and variance σ^2 can be transformed into $\varphi(x)$ through a simple change of variable $x = (y-\mu)/\sigma$. Accordingly, the cumulative integral,

$$P_G(y|\mu,\sigma^2) = \int_{-\infty}^{y} p_G(y'|\mu,\sigma^2)\,dy', \tag{3.135}$$

$$= \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right], \tag{3.136}$$

is equal to $\Phi(x)$.

Returning to the standard normal distributions, $\varphi(x)$, it is useful to consider the probability $p_{DS}(\Delta)$ that the variable x will be found within the double-sided (DS) interval $-\Delta \le x \le \Delta$. It is calculated according to

$$p_{DS}(\Delta) \equiv \frac{1}{\sqrt{2\pi}}\int_{-\Delta}^{\Delta} \exp(-t^2/2)\,dt, \tag{3.137}$$

$$= \Phi(\Delta) - \Phi(-\Delta), \tag{3.138}$$

$$= \text{erf}(\Delta/\sqrt{2}), \tag{3.139}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

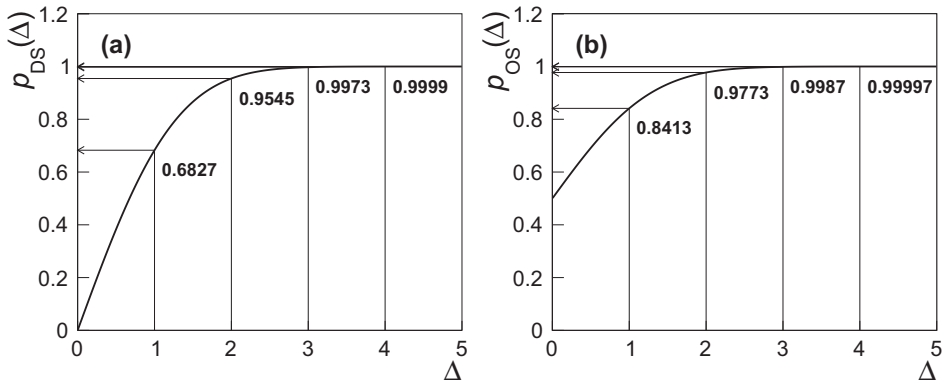


Fig. 3.11 Probability that a random variable x distributed according to a standard normal distribution is found within (a) $\pm\Delta$ of the origin and (b) in the interval $x \leq \Delta$.

shown in Figure 3.11a. The probability that x falls within one "sigma" ($\pm\sigma$) of the mean is 0.6827 and amounts to 0.9545, 0.9973, and 0.999937 for $\Delta = 2\sigma$, 3σ, and 4σ, respectively. One-sided intervals are also of interest when a variable is found at or near a physical limit and one wishes to know the probability of $x \leq \Delta$. One must then consider the one-sided probability $p_{\mathrm{OS}}(\Delta)$ defined by

$$p_{\mathrm{OS}}(\Delta) \equiv \Phi(\Delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\Delta} \exp(-t^2/2)\, dt \tag{3.140}$$

$$= \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\Delta/\sqrt{2}\right), \tag{3.141}$$

shown in Figure 3.11b.

The Gaussian distribution has applications in very many practical situations. This stems largely from the central limit theorem (discussed in §2.10.4), which states that the sum of n independent random numbers x_i, $i = 1, \ldots, n$ of means μ_i and variances σ_i^2 becomes a Gaussian distribution with mean $\mu = \sum_{i=1}^{n} \mu_i$ and variance $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$ in the large n limit. The theorem applies under general conditions regardless of the individuals PDFs of the variables x_i. We will discuss several applications of the central limit theorem and the Gaussian distribution throughout this book.

3.10 Multidimensional Gaussian Distribution

The **multidimensional Gaussian distribution** finds applications in systems involving simultaneously several variables with Gaussian distributions. For instance, consider a system involving m variables, $\vec{x} = (x_1, \ldots, x_m)$ of mean value $\vec{\mu} = (\mu_1, \ldots, \mu_i)$. The multidimensional Gaussian distribution may be written in its most general form:

$$p_{\mathrm{MG}}(\vec{x}) = K \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T \mathbf{Q}\,(\vec{x} - \vec{\mu})\right], \tag{3.142}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

where K is a normalization constant, which we evaluate in the following, and $\mathbf{Q}$ is an $m \times m$ symmetric matrix. There exists a transformation

$$\vec{y} = \mathbf{A}\vec{x} \tag{3.143}$$

such that $\mathbf{Q}$ can be diagonalized. The matrix $\mathbf{A}$ is orthogonal, satisfies $\mathbf{A}^{-1} = \mathbf{A}^T$, and has a determinant $|\mathbf{A}| = 1$. It is then possible to reduce Eq. (3.142) to a simpler form involving a diagonal matrix $\mathbf{D} = \mathbf{AQA}^T$. Defining $\vec{v} = \mathbf{A}\vec{\mu}$, Eq. (3.142) may then be written

$$p_{\mathrm{MG}}(\vec{x}) = K \exp\left[-\frac{1}{2}(\vec{y} - \vec{v})^T \mathbf{AQA}^T (\vec{y} - \vec{v})\right], \tag{3.144}$$

$$= K \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{m} D_{ii}(y_i - v_i)^2\right)\right]. \tag{3.145}$$

The elements D_{ii} can readily be identified as the multiplicative inverse of the variances, $\sigma_{y,i}^{-2}$, of the variables y_i. The normalization constant K is thus equal to $(2\pi)^{-m/2} \prod \sigma_{y,i}^{-1} = (2\pi)^{-m/2}|D|^{-1/2}$. By extension, one also concludes that the matrix $\mathbf{Q}$ corresponds to the inverse, $\mathbf{V}^{-1}$, of the covariance matrix of variables x_i. By construction, one can thus write

$$\mathbf{V} = \mathbf{A}^T \mathbf{D}^{-1} \mathbf{A}, \tag{3.146}$$

and the multi-Gaussian distribution becomes

$$p_{\mathrm{MG}}(\vec{x}) = \frac{1}{(2\pi)^{m/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T \mathbf{V}^{-1} (\vec{x} - \vec{\mu})\right], \tag{3.147}$$

where we also expressed the constant K in terms of the determinant of the covariance matrix $\mathbf{V}$.

As an example of application of the multidimensional Gaussian distribution, consider a Gaussian two-particle correlation function in pseudorapidity.[4] Let η_1 and η_2 represent the pseudorapidity of the two particles. Their relative and average rapidities are written $\Delta\eta = \eta_1 - \eta_2$ and $\bar{\eta} = 0.5(\eta_1 + \eta_2)$, respectively. Let us model the correlation strength between the two particles according to their relative and average rapidities using Gaussian distributions:

$$C(\Delta\eta, \bar{\eta}) \propto \exp\left(-\frac{\Delta\eta^2}{2\sigma_{\Delta\eta}^2}\right) \exp\left(-\frac{\bar{\eta}^2}{2\sigma_{\bar{\eta}}^2}\right), \tag{3.148}$$

where $\sigma_{\Delta\eta}$ and $\sigma_{\bar{\eta}}$ are the widths of the correlations in $\Delta\eta$ and $\bar{\eta}$, respectively. Let us define a matrix D as follows:

$$D = \begin{pmatrix} \sigma_{\Delta\eta}^{-2} & 0 \\ 0 & \sigma_{\bar{\eta}}^{-2} \end{pmatrix}. \tag{3.149}$$

[4] The notion of correlation function will be defined in Chapter 10. However, all that matters here is that the function considered depends on two Gaussian random variables.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Let us also define a matrix A to transform coordinates η_1 and η_2 into $\Delta\eta$ and $\bar{\eta}$ according to

$$\begin{pmatrix} \Delta\eta \\ \bar{\eta} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}. \tag{3.150}$$

The preceding correlation function may then be written in terms of η_1 and η_2 as follows:

$$C(\eta_1, \eta_2) = \frac{1}{2\pi\sigma_{\Delta\eta}\sigma_{\bar{\eta}}} \exp\left(-\frac{1}{2}\vec{\eta}^T Q\vec{\eta}\right), \tag{3.151}$$

with

$$\vec{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \tag{3.152}$$

and

$$Q = A^T DA = \begin{pmatrix} \sigma_{\Delta\eta}^{-2} + \frac{1}{4}\sigma_{\bar{\eta}}^{-2} & -\sigma_{\Delta\eta}^{-2} + \frac{1}{4}\sigma_{\bar{\eta}}^{-2} \\ -\sigma_{\Delta\eta}^{-2} + \frac{1}{4}\sigma_{\bar{\eta}}^{-2} & \sigma_{\Delta\eta}^{-2} + \frac{1}{4}\sigma_{\bar{\eta}}^{-2} \end{pmatrix}, \tag{3.153}$$

which is the inverse of the covariance matrix of η_1 and η_2.

3.11 Log-Normal Distribution

3.11.1 Definition

The **log-normal distribution** is defined as

$$p_{\ln}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \qquad x > 0, \tag{3.154}$$

where μ and σ^2 are the mean and variance, respectively, of the distribution of the random variable $y = \ln x$ which, by construction, has a normal distribution. The cumulative distribution function is obtained by integration:

$$P_{\ln}(x|\mu, \sigma) = \frac{1}{2}\text{erfc}\left(-\frac{\ln x - \mu}{\sqrt{2}\sigma}\right) = \Phi\left(-\frac{\ln x - \mu}{\sigma}\right), \tag{3.155}$$

where erfc and Φ are the complementary error function, and the standard normal CDF, respectively.

3.11.2 Properties of the Log-Normal Distribution

The variables μ and σ^2 are clearly not the mean and variance of the log-normal distribution. Indeed, direct calculation shows (see Problem 3.23) the mean and variance of $p_{\ln}(x|\mu, \sigma^2)$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

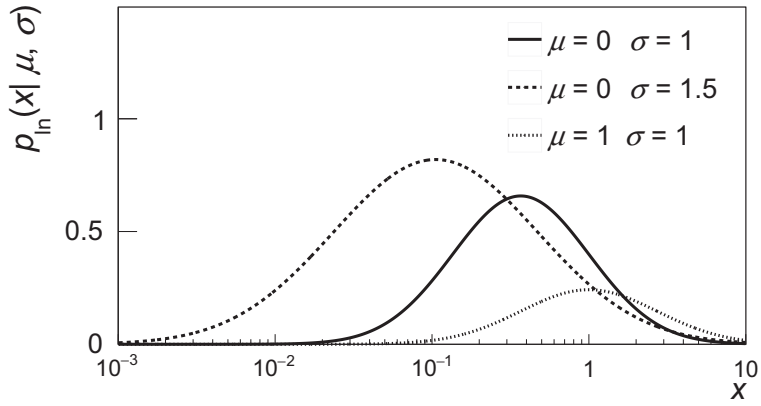


Fig. 3.12 Log-normal distribution $p_{\ln}(x;\,\mu,\,\sigma)$ with selected values of μ and σ.

are as follows:

$$\mathrm{E}[x] = \exp\left(\mu + \frac{1}{2}\sigma^2\right), \tag{3.156}$$

$$\mathrm{Var}[x] = \left[\exp(\sigma^2) - 1\right]\exp\left(2\mu + \sigma^2\right). \tag{3.157}$$

Other moments and properties of the log-normal distribution are presented in Table 3.10. Examples of the log-normal distribution are shown in Figure 3.12.

3.11.3 Generalized Central Limit Theorem

The log-normal distribution is of particular interest with regards to random variables that are the product of many random variables. For instance, consider n random variables z_1, $z_2, \ldots, z_n$ and their product:

$$x \equiv \prod_{i=1}^{n} z_i. \tag{3.158}$$

We introduce variables w_i defined such that $z_i = e^{w_i}$, and $y \equiv \sum_{i=1}^{n} w_i$. By construction, one has $x = e^y$, and the probability $p(x)\,dx$ equals $p(y)\,dy$. This can also be written

$$p(x) \equiv p(y)\left|\frac{dy}{dx}\right|. \tag{3.159}$$

For a reasonably large n, one can apply the central limit theorem and conclude y is Gaussian distributed with mean $\bar{y} \equiv \sum_{i=1}^{n} \bar{w}_i$ and variance $\sigma^2 \equiv \sum_{i=1}^{n} \sigma_{w,i}^2$. The variable x consequently has a log-normal distribution. Indeed, given $\frac{dy}{dx} = \frac{1}{x}$, one obtains:

$$p(x) \equiv \frac{1}{\sqrt{2\pi}\,\sigma}\frac{1}{x}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right). \tag{3.160}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.10 Properties of the log-normal distribution

Function	$p_{\ln}(x;\mu,\sigma) = (2\pi)^{-1/2}(\sigma x)^{-1}\exp\left(-(\ln x-\mu)^2/2\sigma^2\right)$
Characteristic Fct	$\phi(t) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}((\ln x-\mu)/\sqrt{2\sigma^2})$
Mode	$\exp(\mu-\sigma^2)$
Median	$\exp(\mu)$
Mean	$\mu = \exp\left(\mu+\frac{1}{2}\sigma^2\right)$
Variance	$\sigma^2 = (\exp(\sigma^2)-1)\exp(2\mu+\sigma^2)$
Skewness excess	$\gamma_1 = \left(\exp(\sigma^2)+2\right)\sqrt{\exp\sigma^2-1}$
Kurtosis excess	$\gamma_2 = \exp(4\sigma^2)+2\exp(3\sigma^2)+3\exp(2\sigma^2)-6$

The central limit theorem (§2.10.4) can then be "generalized" as follows:

1. The sum of a large number of random variables has a normal distribution.
2. The product of a large number of random variables has a log-normal distribution.

The generalized central limit theorem is applicable, in particular, to a succession of n gain (or attenuation) devices with gains (attenuation factors) $z_i = e^{a_i}$, where the coefficients a_i, which determine the gains, are approximately normally distributed.

The characteristic function and lowest moments of the log-normal distribution are presented in Table 3.10.

3.12 Student's *t*-Distribution

3.12.1 Definition

Student's *t*-distribution, commonly known as the *t-distribution*, involves a family of continuous probability distributions that arise when estimating the average of populations known to have a normal distribution.

Suppose a random variable, X, is known to be normally distributed but with unknown mean and variance. One can then proceed to estimate the mean of the distribution based on a sample consisting of n measurements. As we shall see in Chapter 4, the mean of this sample provides an unbiased estimator of the mean of the distribution. However, the sample mean is expected to deviate, sample by sample, from the actual mean of the distribution. We will show in §6.4 that the probability distribution of the sample mean is given by the t-distribution:

$$p_t(t|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{3.161}$$

where $\nu = n-1$ is the number of degrees of freedom being sampled.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.11 Properties of Student's t-distribution

Function	$p_t(\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
Characteristic Fct	$\phi(t) = \frac{K_{\nu/2}\left(\sqrt{\nu}\lvert t\rvert\right)(1+\nu/2)}{\Gamma(\nu/2)2(\nu/2-1)}$ for $\nu > 0$
	$K_{\nu/2}(x)$ is a Bessel function
Moment Gen Fct	Undefined
Mode	0
Median	0
Mean	$\mu = 0$
Variance	$\sigma^2 = \frac{\nu}{\nu-2}$ for $\nu > 2$
Skewness excess	$\gamma_1 = 0$
Kurtosis excess	$\gamma_2 = \frac{6}{\nu-4}$ for $\nu > 4$

The t-distribution is particularly useful in the context of Student's t-test, which quantifies the likelihood that a measured value lies within a given confidence interval of a predicted value or the results of other measurements (discussed in §6.4). It is symmetric about the origin, $t = 0$, and its moments, relatively simple to calculate, are listed in Table 3.11.

Figure 3.13 presents examples of the t-distribution for selected values of the number of degrees of freedom, ν, compared to a standard normal distribution (thin solid line). For small values of ν, the t-distribution has tails that slowly taper off to zero at infinity. The distribution, however, eventually converges to the normal distribution for a large (infinite) number of degrees of freedom, that is, when the mean is evaluated from very large samples.

3.12.2 Properties of Student's t-Distribution

Properties of the Student's t-distribution are presented in Table 3.11.

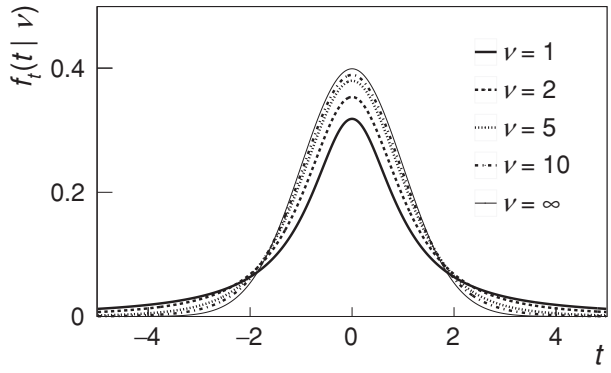


Fig. 3.13 Student's t-distribution for selected values of the number of degrees of freedom, ν, compared to the standard normal distribution (thin solid line).

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

3.13 Chi-Square Distribution

3.13.1 Definition

The χ^2 **distribution** arises in the context of data modeling with the maximum likelihood and least-squares methods discussed in §5.2. We will see that given a dataset consisting of n points (x_i, y_i), $i = 1, \ldots, n$, and a model $f(x|\vec{\theta})$, dependent on m unknown parameters $\vec{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$, one can evaluate the goodness of a fit based on the χ^2 function defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - f(x_i|\vec{\theta})\right]^2}{\sigma_i^2}, \tag{3.162}$$

where σ_i^2 are the errors on the n measurements y_i. For measurements governed by Gaussian statistics, that is, when repeated measurements at a given value of x yield values of y distributed according to a normal distribution, the PDF of the χ^2 values is given by the χ^2-distribution defined as

$$p_{\chi^2}(z|n) \equiv \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} \qquad \text{for } n = 1, 2, \ldots. \tag{3.163}$$

We will derive this distribution in detail in §3.13.3. In this section, let us focus on its properties and interpretation. We first note that the χ^2 distribution of the continuous variable z evaluated in the domain $0 \leq z < \infty$ is a special case of the gamma distribution (§3.6) with $\beta = 1/2$ and $\alpha = n/2$. In this context, the variable n is an integer called **number of degrees of freedom**. The evaluation of the gamma function, $\Gamma(x)$, is thus restricted to x taking integer and half-integer values only. For integers $n \geq 1$, one has $\Gamma(n) = (n-1)!$ whereas for odd values of n, the evaluation of $\Gamma(n/2)$ is achieved with the recursive formula

$$\Gamma(x+1) = x\Gamma(x) \qquad \text{with } \Gamma(1/2) = \sqrt{\pi}. \tag{3.164}$$

The number of degrees of freedom n is a measure of the number of independent variables in a system or model. By definition, an unconstrained dataset containing n independent random variables has n degrees of freedom. However, imagine fixing the mean of the n values y_i prior to their measurements. If the $n-1$ first values are obtained randomly and independently, the nth value becomes automatically constrained by the a priori fixed value of the mean: it cannot be considered random. The dataset has consequently only $n-1$ degrees of freedom. Now, imagine fixing the second moment of the distribution as well. It is easy to verify that once the $n-2$ first measured values have been obtained, the last two are automatically determined by the first $n-2$ values and the specified values of the first and second moments of the distribution. Specifying additional moments (or other linearly independent parameters of the model) thus reduces the number of degrees of freedom: each additional model parameter reduces the number of degrees of freedom by one unit. The number of degrees of freedom of a dataset consisting of n random values fitted to a model involving m parameters is thus $N - m$.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.12 Properties of the χ^2 distribution

Function	$p_{\chi^2}(z;n) \equiv \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$
Characteristic function	$(1-2it)^{n/2}$
Mode	$\max(n-2, 0)$
Median	$\approx n(1-2/9n)^3$
Mean	$\mu = n$
Variance	$\sigma^2 = 2n$
Skewness excess	$\gamma_1 = \sqrt{8/n}$
Kurtosis excess	$\gamma_2 = 12/n$

3.13.2 Properties of the Chi-square Distribution

The mean and variance of the χ^2 distribution may be calculated by integration by parts (see Problem 3.18). One gets

$$\mathrm{E}[z] = \int_0^\infty z \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}\, dz = n, \tag{3.165}$$

and

$$\mathrm{Var}[z] = \int_0^\infty (z-n)^2 \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}\, dz = 2n. \tag{3.166}$$

Other properties of the χ^2 distribution are listed in Table 3.12.

The χ^2 distribution is illustrated in Figure 3.14a for various values of the number of degrees of freedom n. One observes that all distributions exhibit a peak near n. As shown in Figure 3.14b, this translates in having a value of the χ^2 per degrees of freedom,

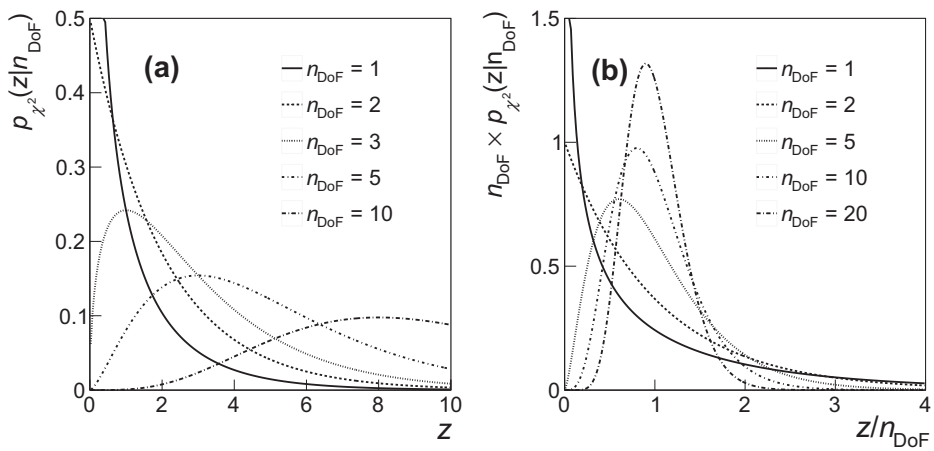


Fig. 3.14 (a) χ^2 probability distribution for selected numbers of degrees of freedom n_{DoF}; (b) scaled χ^2 probability distribution plotted as a function of the χ^2 per degrees of freedom, χ^2/n_{DoF}, for selected values of n_{DoF}.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

$\chi^2/\text{DoF} = \chi^2/n$, peaked near unity. Indeed, note that while the mean of the χ^2/DoF distribution is unity for all values of n, the mode of the distribution differs from unity, with the largest deviations for a small number of degrees of freedom, n.

It is useful to consider the addition of two independent, χ^2-distributed, random variables Z_1 and Z_2 with n_1 and n_2 degrees of freedom, respectively. We will show in §3.13.4 that the sum $Z = Z_1 + Z_2$ is characterized by a χ^2 distribution with $n_1 + n_2$ degrees of freedom. This implies that a merged dataset consisting of two independent datasets behaves as one would intuitively expect, that is, the variable Z shall have a χ^2 distribution determined by the total number of degrees of freedom of the two sets. However, this conclusion does not apply if the two datasets are not mutually independent. This is the case, in particular, if a subset of the first dataset is also included in the second dataset.

3.13.3 Derivation of the χ^2 Distribution

In order to derive the expression of the PDF $p_{\chi^2}(z|n)$ of a χ^2 variable with n degrees of freedom, we appeal to the addition theorem, Eqs. (2.191, 2.193), discussed in §3.13. We define z as the sum of the n error weighted deviations of the measurements y_i, with error σ_i, from the expected values μ_i:

$$z = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2} = \sum_{i=1}^{n} w_i, \tag{3.167}$$

with $w_i = (y_i - \mu_i)^2 / \sigma_i^2$. By virtue of Eq. (2.191), we can write the characteristic function of z as a product of the characteristic functions of the variables w_i.

$$\phi_z(t) = \prod_{i=1}^{n} \phi_{w,i}(t). \tag{3.168}$$

We thus need to determine the function $\phi_{w,i}(t)$. To accomplish this, we first define variables $y_i' = (y_i - \mu_i)/\sigma_i$ and note that since we assumed the variables y_i have normal distributions with mean μ_i and width σ_i, then the variables y_i' follow the standard normal distribution

$$f(y_i') = \frac{1}{\sqrt{2\pi}} e^{-y_i'^2}. \tag{3.169}$$

The PDFs of variables w_i may then be written

$$p_w(w_i) = 2 f(y_i') \left| \frac{dy_i'}{dw_i} \right| = \frac{1}{\sqrt{2\pi\, w_i}} e^{-w_i/2}, \tag{3.170}$$

where the factor of 2 accounts for the fact that two values of y_i' yield the same value w_i. One can show (see Problem 2.23) that the characteristic function of this expression is

$$\phi_{w_i}(t) = (1 - 2it)^{-1/2}. \tag{3.171}$$

According to Eq. (3.168), the characteristic function of z is then given by

$$\phi_z(t) = (1 - 2it)^{-n/2}. \tag{3.172}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Finally, the inverse Fourier transform of this expression yields the χ^2 distribution with n degrees of freedom (see Problem 3.19)

$$p_{\chi^2}(z|n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}. \tag{3.173}$$

3.13.4 Combining the χ^2 of Datasets

The χ^2 distribution is used in a variety of analyses including the evaluation of the goodness-of-fit (§5.3) and tests of compatibility of two independent data samples (§6.6.5). Additionally, since it is natural to expect that a greater amount of data should lead to better constraints on model parameters, it is important to consider how the combination of two independent datasets characterized by χ^2 distributed random variables Z_1 and Z_2 with n_1 and n_2 degrees of freedom, respectively, should behave.

Since Z_1 and Z_2 are assumed to originate from independent datasets, they can be considered independent variables and the characteristic function of their sum Z may then be written

$$\begin{aligned} \phi_Z(t) &= \phi_{Z_1}(t) \times \phi_{Z_1}(t) \\ &= (1-2it)^{-n_1/2} \times (1-2it)^{-n_2/2} \\ &= (1-2it)^{-(n_1+n_2)/2}, \end{aligned} \tag{3.174}$$

where the last line corresponds to the characteristic function of a χ^2 distributed variable with $n_1 + n_2$ degrees of freedom. We thus conclude that the sum Z of two independent χ^2 distributed random variables Z_1 and Z_2 is indeed also a χ^2 distributed random variable with $n_1 + n_2$ degrees of freedom.

3.14 *F*-Distribution

The F-distribution, also known as Snedecor's F-distribution or the Fisher–Snedecor distribution,[5] arises in analyses in which one wishes to establish whether two (or more) datasets have significantly different variances. Such situations commonly occur in manufacturing and quality control: all manufacturing processes lead to finite variance of the quality (specifications) of products. If and when manufacturers change components of their manufacturing process, they may worry that the variance of some particular product specifications (e.g., concentration of medicine in manufactured pills, strength of steel beams used in construction, etc.) may have increased, thereby leading to more product recalls and loss of profits. Similarly, particle detector components built and operated by different agencies or

[5] So named after Ronald Fisher (1890–1962), English statistician, mathematician, and biologist, and George W. Snedecor (1881–1974), American statistician and mathematician.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

technicians may also have different performance variance and consequently impact resolution and detection efficiency differently (e.g., energy resolution of electromagnetic or hadronic calorimeters, §12.2). The F-test is specifically designed for this purpose.

The F-test is formulated on the basis of two independent, χ^2 distributed, random variables Y_1 and Y_2 with n_1 and n_2 degrees of freedom, respectively. One defines the variable f as the ratio of the χ^2 per degree of freedom of two samples:

$$f = \frac{Y_1/n_1}{Y_2/n_2}. \tag{3.175}$$

It can be shown ([5] and references therein) that the random variable f is distributed according to the PDF

$$p(f|n_1, n_2) = \begin{cases} \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{f^{(n_1-2)/2}}{[1+fn_1/n_2]^{(n_1+n_2)/2}} & f > 0 \\ 0 & \text{elsewhere,} \end{cases} \tag{3.176}$$

known as the F-distribution. One additionally finds that the expectation value of f is

$$\langle f \rangle = \frac{n_2}{n_2 - 2}, \tag{3.177}$$

for $n_2 > 2$. Remarkably, this expression is independent of n_1. The variance depends on both n_1 and n_2, however, and one has [5]:

$$\text{Var}[f] = \frac{n_2^2(2n_2 + 2n_1 - 4)}{n_1(n_2 - 1)^2(n_2 - 4)} \tag{3.178}$$

for $n_2 > 4$.

3.15 Breit–Wigner (Cauchy) Distribution

3.15.1 Definition

The **Breit–Wigner** (BW) distribution is named after Gregory Breit[6] (1899–1981) and Eugene Paul Wigner[7] (1902–1995), who contributed to its development and the elucidation of its properties. It is a continuous distribution defined over the domain $-\infty < x < \infty$ as follows:

$$p_{\text{BW}}(x|x_0, \Gamma) = \frac{1}{\pi\Gamma/2} \left[\frac{\Gamma^2/4}{\Gamma^2/4 + (x - x_0)^2}\right]. \tag{3.179}$$

The parameters x_0 and Γ characterize the position (centroid) and width of the distribution, respectively. A special case of the BW distribution, with $x_0 = 0$ and $\Gamma = 2$, is the Cauchy

[6] Russian-born American physicist famous for his work on particle resonant states carried out in collaboration with Wigner, and research with Edward Condon on proton–proton dispersion.

[7] Hungarian-America physicist and mathematician who shared the Nobel Prize in Physics in 1963 for his work on the theory of the atomic nucleus and elementary particles.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

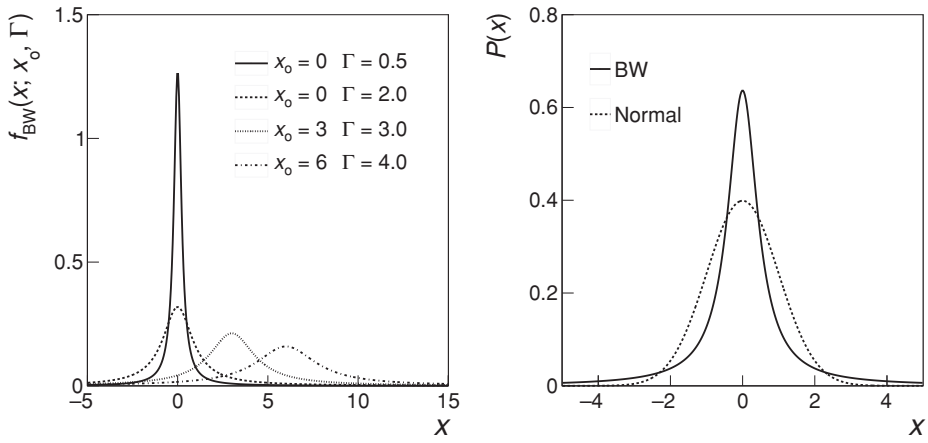


Fig. 3.15 (Left) Breit–Wigner distribution, $p_{BW}(x|x_0, \Gamma)$, for selected values of x_0 and Γ. (Right) comparison of the Breit–Wigner (solid line) and normal (dashed line) distributions for $\Gamma = \sigma = 1.0$.

distribution:

$$p_C(x) = \frac{1}{\pi\,(1 + x^2)}. \tag{3.180}$$

The BW has a particular importance in the calculation of resonant nuclear scattering cross-sections, in which the parameters x_0 and Γ correspond to the mass and width of the resonance, respectively. As such, it finds many applications in experimental data analysis of short-lived particles (e.g., see §8.5.1).

Figure 3.15 (left) illustrates the BW distribution for selected values of the parameters x_o and Γ. The BW distribution may at first sight appear similar in shape to the Gaussian distribution, even though its mathematical expression is obviously quite different. It has, however, rather different properties. Figure 3.15 (right) displays BW and Gaussian distributions centered at the origin ($x_0 = \mu = 0$), and with equal widths ($\Gamma = \sigma$). Indeed, one finds that the two distributions are similar, as they both peak at the origin and taper off asymptotically to zero at infinity. Note, however, that the BW distribution exhibits a much narrower peak and tapers off more slowly than the Gaussian distribution.

In fact, the BW distribution decreases toward zero so slowly that the integrals $\int_0^\infty x f(x)\,dx$ and $\int_{-\infty}^0 x f(x)\,dx$ are individually divergent (see Problem 3.25). This implies that the expectation value E[x] and all higher moments of the distributions are also divergent. That said, by virtue of the symmetry of the distribution about x_0, this value obviously corresponds to the median and can be used to characterize the centroid of the distribution. Likewise, while the parameter Γ cannot be identified as a moment of the distribution, it is nonetheless useful to characterize its width.

3.15.2 Relativistic Breit–Wigner Distribution

In the context of high-energy collision studies where the energy and momenta of particles can be rather large, one modifies the foregoing Breit–Wigner distribution into a Relativistic

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Breit–Wigner (RBW) distribution written

$$p_{\mathrm{RBW}}(E|M,\Gamma) = \frac{k}{(E^2 - M^2)^2 + M^2\Gamma^2}, \tag{3.181}$$

where E, M, Γ are the center-of-mass energy that produces the resonance, the mass of the resonance, and its decay width, respectively. The constant k is given by

$$k = \frac{2\sqrt{2}M\Gamma\gamma}{\pi\sqrt{M^2+\gamma}}, \tag{3.182}$$

with

$$\gamma = \sqrt{M^2\,(M^2+\Gamma^2)}, \tag{3.183}$$

written in natural units, $\hbar = c = 1$. In general, the width Γ may be a function of the energy E when it constitutes a large fraction of the mass M of the resonance (e.g., $\rho^0 \to \pi + \pi$).

3.16 Maxwell–Boltzmann Distribution

3.16.1 Definition

The **Maxwell–Boltzmann distribution**, often called the Maxwell speed distribution, is a probability density describing the velocity distribution of molecules of an idealized gas in thermodynamic equilibrium at a specific temperature T. It can also be used to describe the velocity or energy distribution of an idealized hadron (or parton) gas produced in heavy ion collisions, provided these can be considered equilibrated at a specific temperature. First derived by Maxwell in the late nineteenth century, the distribution may be written

$$p_{\mathrm{MB}}(v) = 4\pi\sqrt{\left(\frac{m}{2\pi kT}\right)^3}\,v^2\exp\left(-\frac{mv^2}{2kT}\right), \tag{3.184}$$

where m stands for the mass of the molecules (particles), v their velocity, T the temperature of the gas, and k the Boltzmann constant. The distribution is illustrated in Figure 3.16 for several values of $\sqrt{kT/m}$.

3.16.2 Properties of the MB Distribution

The most probable velocity, v_{mp}, corresponding to the mode of the distribution, is obtained by setting the derivative of the distribution with respect to v equal to zero. One finds

$$v_{\mathrm{mp}} = \sqrt{\frac{2kT}{m}}. \tag{3.185}$$

The mean speed is given by the expectation value $\mathrm{E}[v]$ of the distribution:

$$\langle v\rangle \equiv \mathrm{E}[v] = \int_0^\infty v\,p_{\mathrm{MB}}(v)dv = \sqrt{\frac{8kT}{\pi m}} = \frac{2}{\pi}v_{\mathrm{mp}} \tag{3.186}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

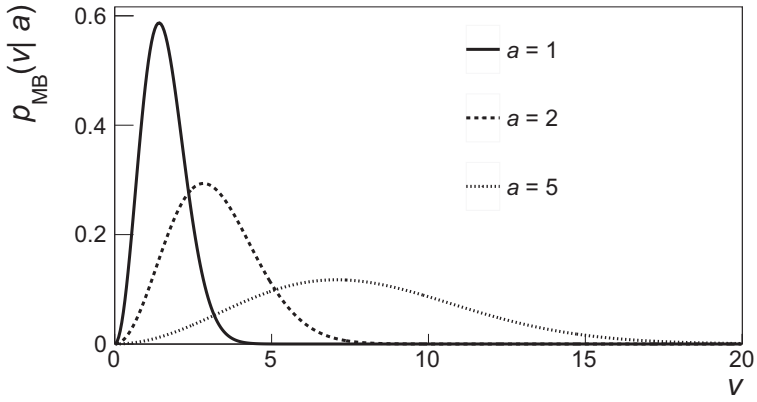


Fig. 3.16 Examples of the Maxwell–Boltzmann distribution for selected values of the parameter $a = \sqrt{kT/m}$.

Usually of greater interest, because it enables the evaluation of the average kinetic energy of the particles, is the root mean square speed, v_{rms}:

$$v_{\mathrm{rms}} \equiv \sqrt{\langle v^2 \rangle} = \left\{ \int_0^\infty v^2 p_{\mathrm{MB}}(v) dv \right\}^{1/2} = \sqrt{\frac{3kT}{m}} = \sqrt{\frac{3}{2}} v_{\mathrm{mp}}. \tag{3.187}$$

3.16.3 Derivation of the MB Distribution

The Maxwell–Boltzmann distribution can be derived using the kinetic theory of gases. In the nonrelativistic limit, the kinetic energy of molecules of type i is

$$E_i = \frac{p_i^2}{2m_i}, \tag{3.188}$$

where p_i and m_i represent the momentum and mass of the molecules (particles), respectively. The number of particles in state i, noted N_i is given by

$$\frac{N_i}{N} = \frac{1}{Z} g_i \exp\left(-\frac{E_i}{kT}\right), \tag{3.189}$$

where T is the equilibrium temperature of the gas, g_i is a degeneracy factor corresponding to the number of microstates that have the same energy levels (or mass), k is the Boltzmann constant, and N is the total number of particles in the system. This can be calculated with the partition function of the system,

$$Z = \sum_j g_j \exp\left(-\frac{E_j}{kT}\right). \tag{3.190}$$

The probability density function, p_{p}, for finding a molecule (particle) at a specific momentum $\vec{p} = (p_x, p_y, p_z)$ can thus be written

$$p_{\mathrm{p}}(p_x, p_y, p_z) = \left(\frac{1}{2\pi mkT}\right)^{3/2} \exp\left(-\frac{p_x^2 + p_y^2 + p_z^2}{2mkT}\right), \tag{3.191}$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

Table 3.13 Properties of the Maxwell–Boltzmann distribution

PDF	$p_{\mathrm{MB}}(x) = \sqrt{\frac{2}{\pi}}\frac{x^2}{a^3}\exp[-x^2/2a^2]$
CDF	$\mathrm{erf}\left(\frac{x}{\sqrt{2}a}\right) - \sqrt{\frac{2}{\pi}}\frac{x\exp(-x^2/(2a^2))}{a}$
Mode	$\sqrt{2}a$
Mean	$\mu = 2^{3/2}\pi^{-1/2}a$
Variance	$\sigma^2 = a^2(3\pi - 8)/\pi$
Skewness	$\gamma_1 = 2^{3/2}(16 - 5\pi)/(3\pi - 8)^{3/2}$
Kurtosis	$\gamma_2 = 4(-96 + 40\pi - 3\pi^2)/(3\pi - 8)^2$

which implies that the momenta p_x, p_y, and p_z are each normally distributed about the origin and with a standard deviation $\sqrt{mkT}$.

The Maxwell–Boltzmann distribution can be expressed in terms of the energy of the particles by a simple variable transformation:

$$p_{\mathrm{E}}(E)\,dE = p_{\mathrm{p}}(\vec{p})\,d^3p = p_{\mathrm{p}}(\vec{p})4\pi p^2 dp. \tag{3.192}$$

Given $dE/dp = p/m$, one gets

$$p_{\mathrm{E}}(E)\,dE = 2\sqrt{\frac{E}{\pi}}\left(\frac{1}{kT}\right)^{3/2}\exp\left(\frac{-E}{kT}\right)dE. \tag{3.193}$$

The properties of the distribution are summarized in Table 3.13.

Exercises

3.1 A vat is filled with 5 yellow, 10 red, 15 green, and 20 blue balls. Calculate the probability of drawing 5 balls of the same color if (a) the balls are put back in the vat after each draw, and (b) if the balls are kept out.

3.2 Calculate the mean, variance, excess skewness, and excess kurtosis of the binomial distribution using a decomposition of $(1-p)^{N-n}$. Repeat these calculations using the moment-generating function of the binomial distribution.

3.3 The designers of a system of interception of ballistic missiles have estimated that their system has an efficiency of 99.8% to intercept any given missile directed at a country. Determine the probability that the system would intercept all of 50 missiles fired against it. How many missiles would an aggressor need to have a probability greater than 0.5 of having one, two, and five missiles hit their target?

3.4 At their peak, typically during August 9–14, the Perseids meteor showers have a rate of 60 meteors per hour, visible from the Northern Hemisphere. Calculate the probability of observing fewer than 10 meteors in 30 minutes, and the probability of observing more than 40 in 10 minutes.

3.5 Use the Stirling formula, Eq. (3.31), to derive the large n limit of the Poisson distribution. Hint: Define $x = \frac{n-\lambda}{\sqrt{\lambda}}$ and use the Stirling formula to express the Poisson distribution in terms of a standard distribution.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

3.6 Measurements at the Intersecting Storage Rings (ISR) and at Fermi National Accelerator Laboratory (FNAL) have shown that the multiplicity of charged particles produced in proton–proton collisions can be described with relatively high precision with a negative binomial distribution. Assuming that all produced particles are pions and neglecting charge conservation effects, derive an expression for the conditional probability of the number of positive pions given fixed numbers of negative and neutral pions. Additionally find an expression for the joint probability of the multiplicity of positive and negative charge particles. Hint: See ref. [197].

3.7 Calculate the mean, variance, and covariance of the multinomial distribution.

3.8 Verify the derivation of Eq. (3.26) and calculate the variance of the multiplicity distribution discussed in §3.1.4.

3.9 The STAR Time Projection Chamber (TPC) detector features 45 sensor rows, thereby enabling reconstruction of charged particles consisting of up to 45 "hits." There is limited efficiency of finding a "hit" on each of the pad rows. Make a plot of the distribution of the number of hits on measured tracks for hit finding efficiencies of 90%, 75%, and 50%. Assume for simplicity's sake that all charged particles cross all 45 pad rows.

3.10 Numerically determine the mean, variance, skewness, and kurtosis of the distributions evaluated in the previous problem. Compare these estimates to those expected from the relationships found in Table 3.1.

3.11 Discuss the trivial distribution of net charge that can be observed for total charged particle multiplicities of 50, 100, 200, and 500 particles. Assume the mean of these net-charge distributions is null. Calculate the variance of each of these distributions.

3.12 Calculate the mean, variance, excess skewness, and excess kurtosis of the Poisson distribution.

3.13 Calculate the mean, variance, excess skewness, and excess kurtosis of the Uniform distribution.

3.14 Calculate the mean, variance, excess skewness, and excess kurtosis of the Exponential distribution. Derive a general expression for all moments and central moments of the Exponential distribution.

3.15 Verify the normalization, Eq. (3.127), of the normal distribution.

3.16 Calculate the mean, variance, excess skewness, and excess kurtosis of the Gaussian distribution. Derive a general expression for all moments and central moments of the Gaussian distribution.

3.17 Calculate the first three cumulants, κ_n, $n = 1, 2, 3$ of the Poisson and Gaussian distributions.

3.18 Calculate the mean, and variance of the χ^2 distribution.

3.19 Calculate the inverse Fourier transform of $\phi_z(t) = (1 - 2it)^{-n/2}$ and derive the expression (3.173) of the χ^2 distribution.

3.20 Calculate the probability that an unstable particle, for example, a lambda baryon, decays within a distance x of its site of production. Assume the particle has a half-life $t_{1/2}$ and neglect relativistic effects.

3.21 Derive a relationship between the mean lifetime of a particle and its half-life.

3.22 Calculate the mean and variance of the χ^2 distribution using integration by parts.

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press

3.23 Derive the expressions Eq. (3.156) for the mean and variance of the log-normal distribution.

3.24 Calculate the probability the χ^2 of a fit (where the χ^2-distribution applies) is larger than the number of degrees of freedom, n, for values of n equal to 2, 4, 8, 16, and 32.

3.25 Show that the integral $\int_0^\infty E^n f_{\mathrm{BW}}(E)\,dE$, where $f_{\mathrm{BW}}(E; x_o, \Gamma)$ is the Breit–Wigner distribution, does not converge for n equal to or larger than 1.

3.26 Implement a computer program to carry out numerical calculations of the cumulative integral of the Gaussian and Breit–Wigner distributions. Use this program to calculate the probability that values lie beyond 1.5 σ, 2.5 σ, and 3.5 σ from the mode of these distributions (Use $\Gamma = \sigma$ for the Breit–Wigner distribution).

3.27 Derive the expression (3.191) for the Maxwell–Boltzmann distribution in terms of the momenta p_x, p_y, and p_z.

3.28 Use Eq. (3.191) to derive the expression of the Maxwell–Boltzmann given by Eq. (3.184).

3.29 Show that in thermodynamic equilibrium, the energy per degree of freedom, ϵ, of the molecules of a gas (of identical particles) is distributed as a chi-squared distribution with one degree of freedom according to

$$f_\epsilon(\epsilon)\,d\epsilon = \sqrt{\frac{1}{\epsilon \pi kT}} \exp\left[\frac{-\epsilon}{kT}\right] d\epsilon.$$

https://doi.org/10.1017/9781108241922.005 Published online by Cambridge University Press