

DRAFT V1.1

From

Math, Numerics, & Programming
(for Mechanical Engineers)

George Konidakis
James Douglass Penn
Masayuki Yano
Anthony T Patera

1 February 2012

© The Authors

License: [Creative Commons Attribution-Noncommercial-Share Alike 3.0](https://creativecommons.org/licenses/by-nc-sa/3.0/) (CC BY-NC-SA 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original authors and MIT OpenCourseWare source are credited; the use is non-commercial; and the CC BY-NC-SA license is retained. See also <http://ocw.mit.edu/terms/>.

Contents

II	Monte Carlo Methods.	115
8	Motivation	117
8.1	Introduction	117
8.2	Occupancy Grid	117
8.3	Monte Carlo Methods	119
9	Introduction to Random Variables	121
9.1	Discrete Random Variables	121
9.1.1	Transformation	127
9.2	Discrete Bivariate Random Variables (Random Vectors)	129
9.2.1	Joint Distributions	129
9.2.2	Characterization of Joint Distributions	130
9.3	Binomial Distribution	136
9.4	Continuous Random Variables	141
9.4.1	Description	141
9.4.2	Transformations of Continuous Random Variables	144
9.4.3	The Central Limit Theorem	148
9.4.4	Generation of Pseudo-Random Numbers	149
9.5	Continuous Random Vectors	150
10	Estimation: Bernoulli Coins	159
10.1	The Sample Mean: An Estimator / Estimate	159
10.2	Confidence Intervals	162
10.2.1	Frequentistic Interpretation	163
10.2.2	Convergence	165
10.3	Cumulative Sample Means	165
11	Estimation: Areas and Volumes	167
11.1	Calculating an Area	167
11.1.1	Objective	167
11.1.2	A Continuous Uniform Random Variable	167
11.1.3	A Bernoulli Random Variable	167
11.1.4	Estimation: Monte Carlo	168
11.1.5	Estimation: Riemann Sum	170
11.2	Calculation of Volumes in Higher Dimensions	173
11.2.1	Three Dimensions	173
	Monte Carlo	173
	Riemann Sum	174
11.2.2	General d -Dimensions	175

12 Integration: General Procedures	179
13 Failure Probabilities	181
13.1 Calculating a Failure Probability	181
13.1.1 Objective	181
13.1.2 An Integral	182
13.1.3 A Monte Carlo Approach	182

Unit II

Monte Carlo Methods.

Chapter 8

Motivation

DRAFT V1.1 © The Authors. License: Creative Commons [BY-NC-SA 3.0](#)

8.1 Introduction

A museum has enlisted a camera-equipped mobile robot for surveillance purposes. The robot will navigate the museum’s premises, pausing to take one or more 360 degree scans in each room. Figure 8.1 shows a typical room filled with various stationary obstructions (in black). We wish to determine the vantage point in each room from which the robot will have the most unobstructed view for its scan by *estimating the visible area (in white) for each candidate vantage point*. We may also wish to provide the robot with an “onboard” visible area estimator for purposes of real-time adaptivity, for example, if the room configuration is temporarily modified.

We first define, for any vantage point \mathbf{x}_V and any surveillance point (to be watched) in the room \mathbf{x}_W , the line segment $S(\mathbf{x}_V, \mathbf{x}_W)$ that connects \mathbf{x}_V and \mathbf{x}_W . We can then express the area visible from a vantage point \mathbf{x}_V as the integral

$$A(\mathbf{x}_V) = \int_{\mathbf{x}_W \in R \text{ such that } S(\mathbf{x}_V, \mathbf{x}_W) \cap O = \emptyset} d\mathbf{x}_W, \quad (8.1)$$

where R is the room and O is the collection of obstructions. The visible area is thus defined as the integral over all points in the room such that the line segment $S(\mathbf{x}_V, \mathbf{x}_W)$ between \mathbf{x}_V and \mathbf{x}_W does not intersect an obstruction (or, equivalently, such that the intersection of sets S and O is the null set).

8.2 Occupancy Grid

There are many ways to do the visibility test $S(\mathbf{x}_V, \mathbf{x}_W) \cap O \stackrel{?}{=} \emptyset$, but perhaps the method most amenable to mobile robotics is to use an “occupancy grid,” a discretization of the map in which each cell’s value corresponds to the likelihood that the cell is empty or occupied. We begin by converting our map of the room to an “occupancy grid,” a discretization of the map in which each cell’s value corresponds to the likelihood that the cell is empty or occupied. In our case, because we know ahead of time the layout of the room, a given cell contains either a zero if the cell is empty, or a one if it is occupied. Figure 8.2(a) shows a visualization of a fairly low-resolution occupancy grid for our map, where occupied cells are shown in black. Figure 8.2(b) shows the discretization convention we have adopted for our grid, in which the cell in row i and column j corresponds to continuous, non-dimensionalized space $i - 1 < x^* \leq i$ and $j - 1 < y^* \leq j$.

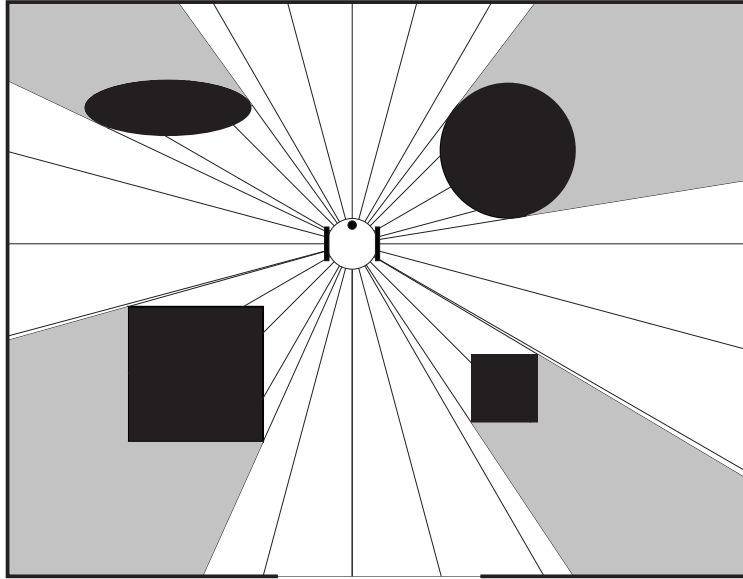


Figure 8.1: A surveillance robot scanning a room. Obstructions (in black) divide the space into visible area (in white) and non-visible area (in gray).

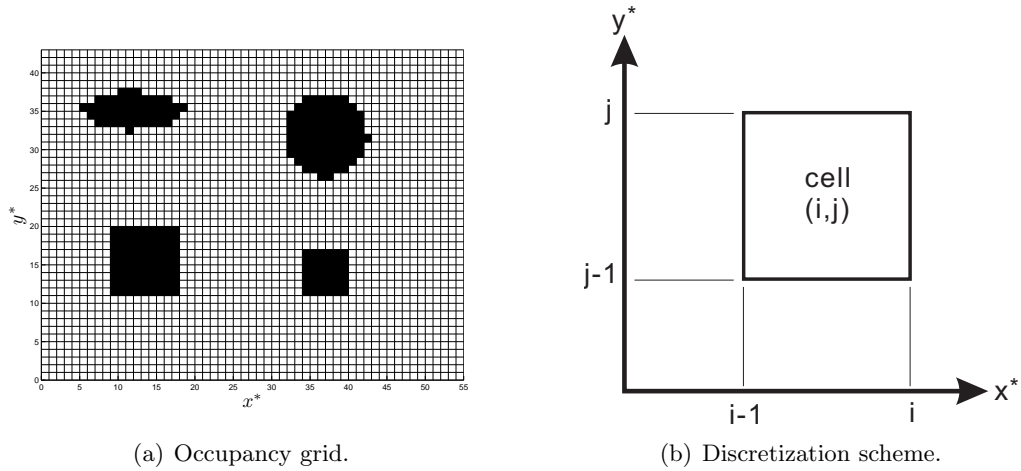


Figure 8.2: Occupancy grid and associated discretization scheme.

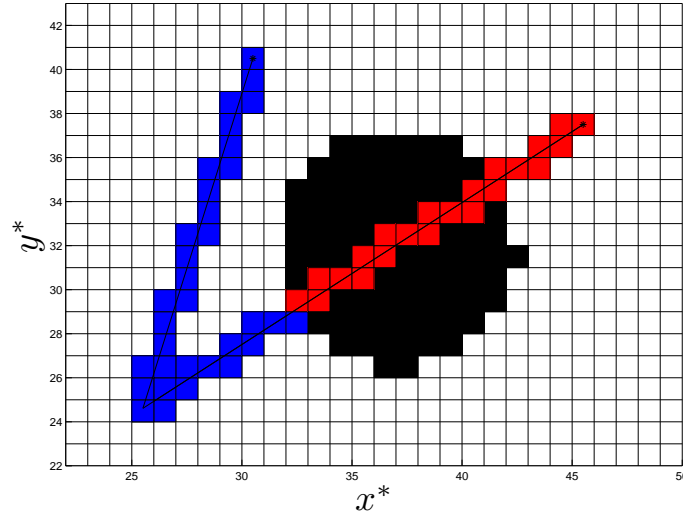


Figure 8.3: Visibility checking of two points from a single vantage point. Visible cells marked in blue, non-visible cells marked in red.

We can use the occupancy grid to determine the visibility of a point \mathbf{x}_W in the room from a given vantage point \mathbf{x}_V . To do this, we draw a line between the two points, determine through which cells the line passes and then check the occupancy condition of each of the intervening cells. If all of the cells are empty, the point is visible. If any of the cells are occupied, the point is not visible. Figure 8.3 shows examples of visible and non-visible cells.

8.3 Monte Carlo Methods

Even once we can evaluate visible or not visible the integral (8.1) remains quite difficult: we have no explicit expression for the domain — which note is not the white space in the room but rather the white space in the room that can be seen by the robot — and we expect that this domain in general will be quite complicated. The deterministic methods discussed in Chapter 7 can indeed be applied. However, Monte Carlo methods have some advantages.

First, Monte Carlo methods are simple to implement: in our case, we do not need to know the domain, we only need to know whether we are in the domain. Second, Monte Carlo methods do not rely on smoothness for convergence — if we think of our integrand as 0 and 1 (depending on visible or not), our problem here is quite non-smooth. Third, although Monte Carlo methods do not converge particularly quickly, the convergence rate does not degrade in higher dimensions — for example, if we wished to construct a 3-D visibility estimate. Fourth, Monte Carlo methods provide a result, along with a simple built-in error estimator, “gradually” — useful, if not particularly accurate, answers are obtained early on in the process and hence inexpensively and quickly. Note for relatively smooth problems in smooth domains Monte Carlo techniques are not a particularly good idea. Different methods work better in different contexts.

Chapter 9

Introduction to Random Variables

DRAFT V1.1 © The Authors. License: Creative Commons [BY-NC-SA 3.0](#)

9.1 Discrete Random Variables

In science and engineering environments, we encounter experiments whose outcome cannot be determined with certainty. While the governing process may be completely deterministic, the uncertainty in the initial condition or the model would render the outcome essentially random. A coin flip and a roll of a die are classical examples of such *random experiments*. The outcome of a random experiment is described by a *random variable* X that takes on a finite number of values,

$$x_1, \dots, x_J ,$$

where J is the number of values that X takes. To fully characterize the behavior of the random variable, we assign a probability to each of these events, i.e.

$$X = x_j, \quad \text{with probability } p_j, \quad j = 1, \dots, J .$$

The same information can be expressed in terms of the *probability mass function* (pmf), or discrete density function, f_X , that assigns a probability to each possible outcome

$$f_X(x_j) = p_j, \quad j = 1, \dots, J .$$

In order for f_X to be a valid probability density function, $\{p_j\}$ must satisfy

$$\begin{aligned} 0 \leq p_j \leq 1, \quad j = 1, \dots, J , \\ \sum_{j=1}^J p_j = 1 . \end{aligned}$$

The first condition requires that the probability of each event be positive and be less than unity. The second condition states that $\{x_1, \dots, x_J\}$ consists the all possible values that X can take, and that the sum of the probabilities of the outcome is unity. The second condition follows from the fact that events x_i , $i = 1, \dots, J$, are *mutually exclusive* and *exhaustive*. Mutually exclusive means that X cannot take on two different values of x_i 's at the same time. The events are exhaustive because X must take on one of the J possible values.

Let us define a few notions useful for characterizing the behavior of the random variable. The expectation of X , $E[X]$, is defined as

$$E[X] = \sum_{j=1}^J x_j p_j .$$

The expectation of X is also called the mean. We denote the mean by μ or μ_X , with the second notation emphasizing that it is the mean of X . Note that the mean is a weighted average of the values taken by X , where each weight is specified according to the respective probability. This is analogous to the concept of moment in mechanics, where the distances are provided by x_j and the weights are provided by p_j ; for this reason, the mean is also called the first moment. The mean corresponds to the centroid in mechanics.

The variance, or the second moment about the mean, measures the spread of the values about the mean and is defined by

$$\text{Var}[X] \equiv E[(X - \mu)^2] = \sum_{j=1}^J (x_j - \mu)^2 p_j .$$

We denote the variance as σ^2 . The variance can also be expressed as

$$\begin{aligned} \text{Var}[X] &= E[(X - \mu)^2] = \sum_{j=1}^J (x_j - \mu)^2 p_j = \sum_{j=1}^J (x_j^2 - 2x_j\mu + \mu^2) p_j \\ &= \underbrace{\sum_{j=1}^J x_j^2 p_j}_{E[X^2]} - 2\mu \underbrace{\sum_{j=1}^J x_j p_j}_{\mu} + \mu^2 \underbrace{\sum_{j=1}^J p_j}_1 = E[X^2] - \mu^2 . \end{aligned}$$

Note that the variance has the unit of X squared. Another useful measure of the expected spread of the random variable is standard deviation, σ , which is defined by

$$\sigma = \sqrt{\text{Var}[X]} .$$

We typically expect departures from the mean of many standard deviations to be rare. This is particularly the case for random variables with large range, i.e. J large. For discrete random variables with small J , this spread interpretation is sometimes not obvious simply because the range of X is small.

The mean and variance (or standard deviation) provide a convenient way of characterizing the behavior of a probability mass function. The two parameters do not provide a full description of the probability distribution, but are often useful (and sometimes sufficient). These parameters are also useful for constructing a model from observed data, where the parameters of the model is tuned to match the statistics of the observation in terms of the mean and variance.

Let us consider a few examples of discrete random variables.

Example 9.1.1 rolling a die

As the first example, let us apply the aforementioned framework to rolling of a die. The random variable X describes the outcome of rolling a (fair) six-sided die. It takes on one of six possible values, $1, 2, \dots, 6$. These events are mutually exclusive, because a die cannot take on two different values at the same time. Also, the events are exhaustive because the die must take on one of the six values after each roll. Thus, the random variable X takes on one of the six possible values,

$$x_1 = 1, x_2 = 2, \dots, x_6 = 6 .$$

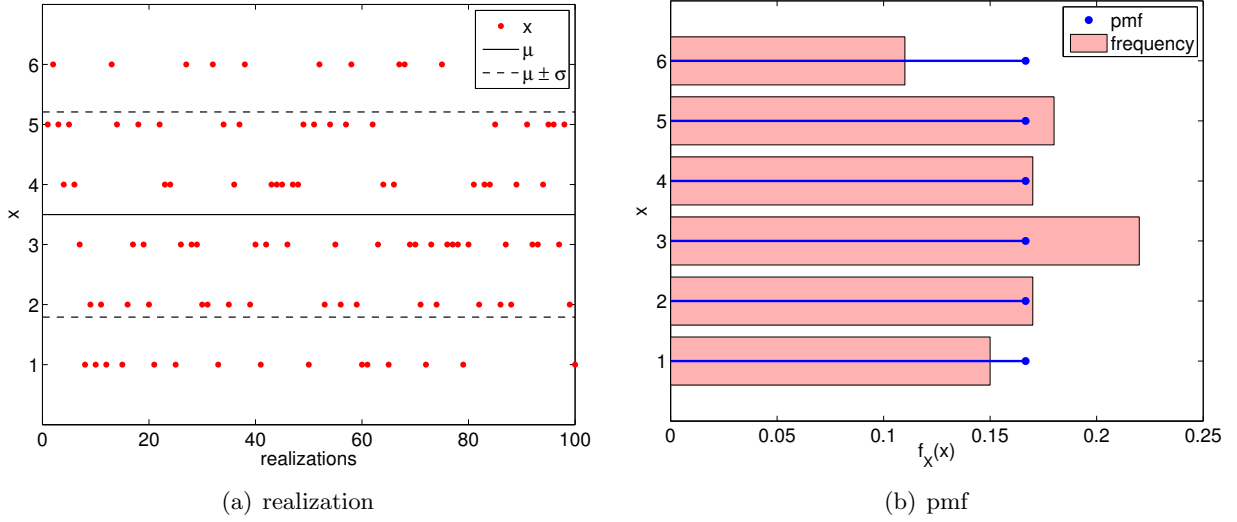


Figure 9.1: Illustration of the values taken by a fair six-sided die and the probability mass function.

A fair die has the equal probability of producing one of the six outcomes, i.e.

$$X = x_j = j, \quad \text{with probability } \frac{1}{6}, \quad j = 1, \dots, 6,$$

or, in terms of the probability mass function,

$$f_X(x) = \frac{1}{6}, \quad x = 1, \dots, 6.$$

An example of outcome of a hundred die rolls is shown in Figure 9.1(a). The die always takes on one of the six values, and there is no obvious inclination toward one value or the other. This is consistent with the fact that any one of the six values is equally likely.

Figure 9.1(b) shows the probability mass function, f_X , of the six equally likely events. The figure also shows the relative frequency of each event — which is defined as the number of occurrences of the event normalized by the total number of samples (which is 100 for this case) — as a histogram. Even for a relatively small sample size of 100, the histogram roughly matches the probability mass function. We can imagine that as the number of samples increases, the relative frequency of each event gets closer and closer to its value of pmf.

Conversely, if we have an experiment with an unknown probability distribution, we could infer its probability distribution through a large number of trials. This follows from the *frequentistic interpretation* of probability: the probability of an event is the relative frequency of its occurrence in a large number of samples. The inference of the underlying probability distribution from a limited amount of data, or samples, is an important problem often encountered in engineering practice.

Let us now characterize the probability mass function in terms of the mean and variance. The mean of the distribution is

$$\mu = E[X] = \sum_{j=1}^6 x_j p_j = \sum_{j=1}^6 j \cdot \frac{1}{6} = \frac{7}{2}.$$

The variance of the distribution is

$$\sigma^2 = \text{Var}[X] = E[X^2] - \mu^2 = \sum_{j=1}^6 x_j^2 p_j - \mu^2 = \sum_{j=1}^6 j^2 \cdot \frac{1}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} \approx 2.9167,$$

and the standard deviation is

$$\sigma = \sqrt{\overline{\text{Var}[X]}} = \sqrt{\frac{35}{12}} = 1.7078 .$$

Example 9.1.2 (discrete) uniform distribution

The outcome of rolling a (fair) die is a special case of a more general distribution, called the (discrete) uniform distribution. The uniform distribution is characterized by each event having the equal probability. It is described by two integer parameters, a and b , which assign the lower and upper bounds of the sample space, respectively. The distribution takes on $J = b - a + 1$ values. For the six-sided die, we have $a = 1$, $b = 6$, and $J = b - a + 1 = 6$. In general, we have

$$x_j = a + j - 1, \quad j = 1, \dots, J ,$$

$$f(x) = \frac{1}{J} .$$

The mean and variance of a (discrete) uniform distribution are given by

$$\mu = \frac{a + b}{2} \quad \text{and} \quad \sigma^2 = \frac{J^2 - 1}{12} .$$

We can easily verify that the expressions are consistent with the die rolling case with $a = 1$, $b = 6$, and $J = 6$, which result in $\mu = 7/2$ and $\sigma^2 = 35/12$.

Proof. The mean follows from

$$\begin{aligned} \mu &= E[X] = E[X - (a - 1) + (a - 1)] = E[X - (a - 1)] + a - 1 \\ &= \sum_{j=1}^J (x_j - (a - 1))p_j + a - 1 = \sum_{j=1}^J j \frac{1}{J} + a - 1 = \frac{1}{J} \frac{J(J + 1)}{2} + a - 1 \\ &= \frac{b - a + 1 + 1}{2} + a - 1 = \frac{b + a}{2} . \end{aligned}$$

The variance follows from

$$\begin{aligned} \sigma^2 &= \text{Var}[X] = E[(X - E[X])^2] = E[((X - (a - 1)) - E[X - (a - 1)])^2] \\ &= E[(X - (a - 1))^2] - E[X - (a - 1)]^2 \\ &= \sum_{j=1}^J (x_j - (a - 1))^2 p_j - \left[\sum_{j=1}^J (x_j - (a - 1)) p_j \right]^2 \\ &= \sum_{j=1}^J j^2 \frac{1}{J} - \left[\sum_{j=1}^J j p_j \right]^2 = \frac{1}{J} \frac{J(J + 1)(2J + 1)}{6} - \left[\frac{1}{J} \frac{J(J + 1)}{2} \right]^2 \\ &= \frac{J^2 - 1}{12} = \frac{(b - a + 1)^2 - 1}{12} . \end{aligned}$$

□

Example 9.1.3 Bernoulli distribution (a coin flip)

Consider a classical random experiment of flipping a coin. The outcome of a coin flip is either a

head or a tail, and each outcome is equally likely assuming the coin is fair (i.e. unbiased). Without loss of generality, we can associate the value of 1 (success) with head and the value of 0 (failure) with tail. In fact, the coin flip is an example of a Bernoulli experiment, whose outcome takes on either 0 or 1.

Specifically, a Bernoulli random variable, X , takes on two values, 0 and 1, i.e. $J = 2$, and

$$x_1 = 0 \quad \text{and} \quad x_2 = 1.$$

The probability mass function is parameterized by a single parameter, $\theta \in [0, 1]$, and is given by

$$f_X(x; \theta) = \begin{cases} 1 - \theta, & x = 0 \\ \theta, & x = 1. \end{cases}$$

In other words, θ is the probability that X takes the value of 1. Flipping of a fair coin is a particular case of a Bernoulli experiment with $\theta = 1/2$. The $\theta = 1/2$ case is also a special case of the discrete uniform distribution with $a = 0$ and $b = 1$, which results in $J = 2$.

Examples of the values taken by Bernoulli random variables ($\theta = 1/2$ and $\theta = 1/4$) are shown in Figure 9.2. As expected, with $\theta = 1/2$, the random variable takes on the value of 0 and 1 roughly the equal number of times. On the other hand, $\theta = 1/4$ results in the random variable taking on 0 more frequently than 1.

The probability mass functions, shown in Figure 9.2, reflect the fact that $\theta = 1/4$ results in X taking on 0 three times more frequently than 1. Even with just 100 samples, the relative frequency histograms captures the difference in the frequency of the events for $\theta = 1/2$ and $\theta = 1/4$. In fact, even if we did not know the underlying pmf — characterized by θ in this case — we can infer from the sampled data that the second case has a lower probability of success (i.e. $x = 1$) than the first case. In the subsequent chapters, we will formalize this notion of inferring the underlying distribution from samples and present a method for performing the task.

The mean and variance of the Bernoulli distribution are given by

$$E[X] = \theta \quad \text{and} \quad \text{Var}[X] = \theta(1 - \theta) .$$

Note that lower θ results in a lower mean, because the distribution is more likely to take on the value of 0 than 1. Note also that the variance is small for either $\theta \rightarrow 0$ or $\theta \rightarrow 1$ as in these cases we are almost sure to get one or the other outcome. But note that (say) $\sigma/E(X)$ scales as $1/\sqrt{\theta}$ (recall σ is the standard deviation) and hence the *relative* variation in X becomes *more* pronounced for small θ : this will have important consequences in our ability to predict rare events.

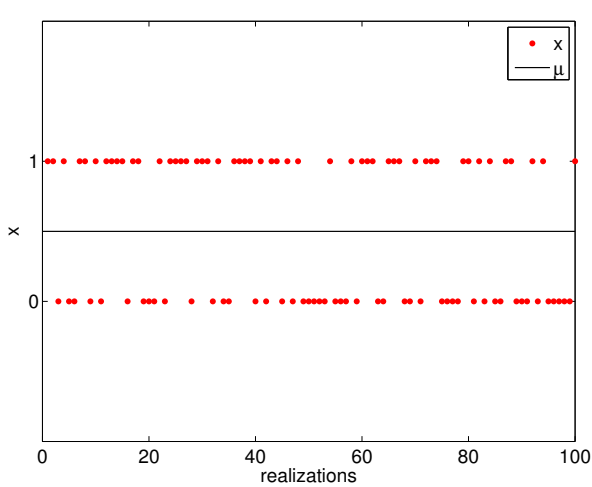
Proof. Proof of the mean and variance follows directly from the definitions. The mean is given by

$$\mu = E[X] = \sum_{j=1}^J x_j p_j = 0 \cdot (1 - \theta) + 1 \cdot \theta = \theta .$$

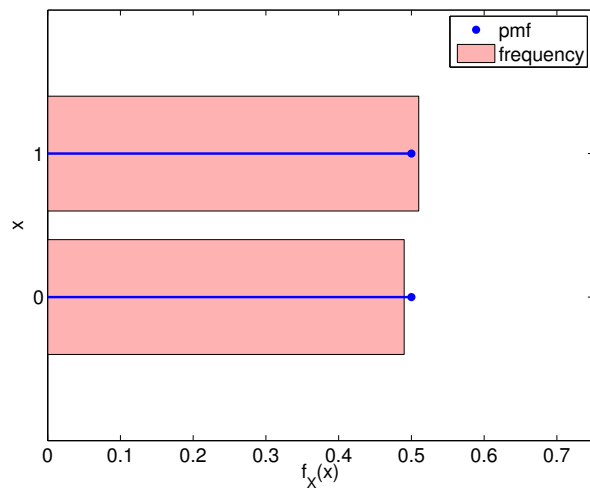
The variance is given by

$$\text{Var}[X] = E[(X - \mu)^2] = \sum_{j=1}^J (x_j - \mu)^2 p_j = (0 - \theta)^2 \cdot (1 - \theta) + (1 - \theta)^2 \cdot \theta = \theta(1 - \theta) .$$

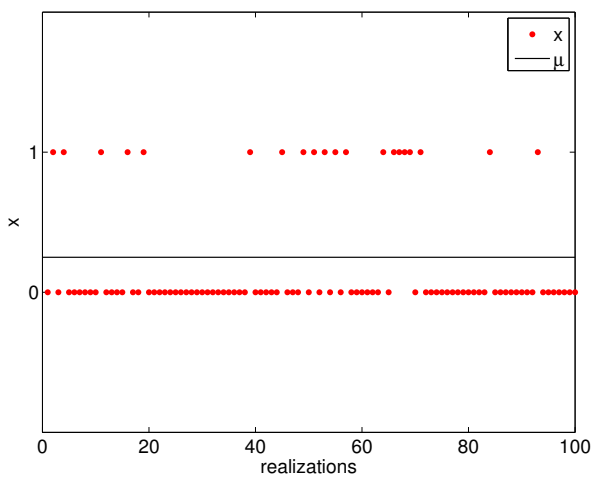
□



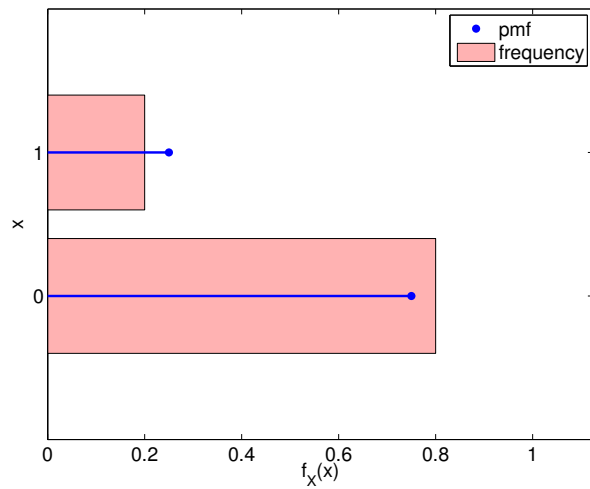
(a) realization, $\theta = 1/2$



(b) pmf, $\theta = 1/2$



(c) realization, $\theta = 1/4$



(d) pmf, $\theta = 1/4$

Figure 9.2: Illustration of the values taken by Bernoulli random variables and the probability mass functions.

9.1.1 Transformation

Random variables, just like deterministic variables, can be transformed by a function. For example, if X is a random variable and g is a function, then a transformation

$$Y = g(X)$$

produces another random variable Y . Recall that we described the behavior of X that takes on one of J values by

$$X = x_j \quad \text{with probability } p_j, \quad j = 1, \dots, J.$$

The associated probability mass function was $f_X(x_j) = p_j$, $j = 1, \dots, J$. The transformation $Y = g(X)$ yields the sample space y_j , $j = 1, \dots, J$, where each y_j results from applying g to x_j , i.e.

$$y_j = g(x_j), \quad j = 1, \dots, J.$$

Thus, Y can be described by

$$Y = y_j = g(x_j) \quad \text{with probability } p_j, \quad j = 1, \dots, J.$$

We can write the probability mass function of Y as

$$f_Y(y_j) = f_Y(g(x_j)) = p_j \quad j = 1, \dots, J.$$

We can express the mean of the transformed variable in a few different ways:

$$E[Y] = \sum_{j=1}^J y_j f_Y(y_j) = \sum_{j=1}^J y_j p_j = \sum_{j=1}^J g(x_j) f_X(x_j).$$

The first expression expresses the mean in terms of Y only, whereas the final expression expresses $E[Y]$ in terms of X and g without making a direct reference to Y .

Let us consider a specific example.

Example 9.1.4 from rolling a die to flipping a coin

Let us say that we want to create a random experiment with equal probability of success and failure (e.g. deciding who goes first in a football game), but all you have is a die instead of a coin. One way to create a Bernoulli random experiment is to roll the die, and assign “success” if an odd number is rolled and assign “failure” if an even number is rolled.

Let us write out the process more formally. We start with a (discrete) uniform random variable X that takes on

$$x_j = j, \quad j = 1, \dots, 6,$$

with probability $p_j = 1/6$, $j = 1, \dots, 6$. Equivalently, the probability density function for X is

$$f_X(x) = \frac{1}{6}, \quad x = 1, 2, \dots, 6.$$

Consider a function

$$g(x) = \begin{cases} 0, & x \in \{1, 3, 5\} \\ 1, & x \in \{2, 4, 6\}. \end{cases}$$

Let us consider a random variable $Y = g(X)$. Mapping the outcomes of X, x_1, \dots, x_6 , to y'_1, \dots, y'_6 , we have

$$\begin{aligned} y'_1 &= g(x_1) = g(1) = 0 , \\ y'_2 &= g(x_2) = g(2) = 1 , \\ y'_3 &= g(x_3) = g(3) = 0 , \\ y'_4 &= g(x_4) = g(4) = 1 , \\ y'_5 &= g(x_5) = g(5) = 0 , \\ y'_6 &= g(x_6) = g(6) = 1 . \end{aligned}$$

We could thus describe the transformed variable Y as

$$Y = y'_j \quad \text{with probability } p_j = 1/6, \quad j = 1, \dots, 6 .$$

However, because $y'_1 = y'_3 = y'_5$ and $y'_2 = y'_4 = y'_6$, we can simplify the expression. Without loss of generality, let us set

$$y_1 = y'_1 = y'_3 = y'_5 = 0 \quad \text{and} \quad y_2 = y'_2 = y'_4 = y'_6 = 1 .$$

We now combine the frequentistic interpretation of probability with the fact that x_1, \dots, x_6 are mutually exclusive. Recall that to a frequentist, $P(Y = y_1 = 0)$ is the probability that Y takes on 0 in a large number of trials. In order for Y to take on 0, we must have $x = 1, 3$, or 5 . Because X taking on 1, 3, and 5 are mutually exclusive events (e.g. X cannot take on 1 and 3 at the same time), the number of occurrences of $y = 0$ is equal to the sum of the number of occurrences of $x = 1$, $x = 3$, and $x = 5$. Thus, the relative frequency of Y taking on 0 — or its probability — is equal to the sum of the relative frequencies of X taking on 1, 3, or 5. Mathematically,

$$P(Y = y_1 = 0) = P(X = 1) + P(X = 3) + P(X = 5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} .$$

Similarly, because X taking on 2, 4, and 6 are mutually exclusive events,

$$P(Y = y_2 = 1) = P(X = 2) + P(X = 4) + P(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} .$$

Thus, we have

$$Y = \begin{array}{ll} 0, & \text{with probability } 1/2 \\ 1, & \text{with probability } 1/2 , \end{array}$$

or, in terms of the probability density function,

$$f_Y(y) = \frac{1}{2}, \quad y = 0, 1 .$$

Note that we have transformed the uniform random variable X by the function g to create a Bernoulli random variable Y . We emphasize that the mutually exclusive property of x_1, \dots, x_6 is the key that enables the simple summation of probability of the events. In essence, (say), $y = 0$ obtains if $x = 1$ OR if $x = 3$ OR if $x = 5$ (a union of events) and since the events are mutually exclusive the “number of events” that satisfy this condition — ultimately (when normalized) frequency or probability — is the sum of the individual “number” of each event. The transformation procedure is illustrated in Figure 9.3.

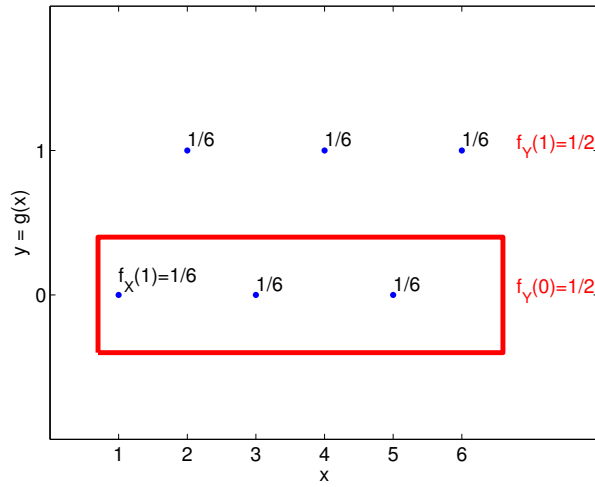


Figure 9.3: Transformation of X associated with a die roll to a Bernoulli random variable Y .

Let us now calculate the mean of Y in two different ways. Using the probability density of Y , we can directly compute the mean as

$$E[Y] = \sum_{j=1}^2 y_j f_Y(y_j) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2} .$$

Or, we can use the distribution of X and the function g to compute the mean

$$E[Y] = \sum_{j=1}^6 g(x_j) f_X(x_j) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} = \frac{1}{2} .$$

Clearly, both methods yield the same mean.

9.2 Discrete Bivariate Random Variables (Random Vectors)

9.2.1 Joint Distributions

So far, we have consider scalar random variables, each of whose outcomes is described by a single value. In this section, we extend the concept to random variables whose outcome are vectors. For simplicity, we consider a random vector of the form

$$(X, Y) ,$$

where X and Y take on J_X and J_Y values, respectively. Thus, the random vector (X, Y) takes on $J = J_X \cdot J_Y$ values. The probability mass function associated with (X, Y) is denoted by $f_{X,Y}$. Similar to the scalar case, the probability mass function assigns a probability to each of the possible outcomes, i.e.

$$f_{X,Y}(x_i, y_j) = p_{ij}, \quad i = 1, \dots, J_X, \quad j = 1, \dots, J_Y .$$

Again, the function must satisfy

$$0 \leq p_{ij} \leq 1, \quad i = 1, \dots, J_X, \quad j = 1, \dots, J_Y ,$$

$$\sum_{j=1}^{J_Y} \sum_{i=1}^{J_X} p_{ij} = 1 .$$

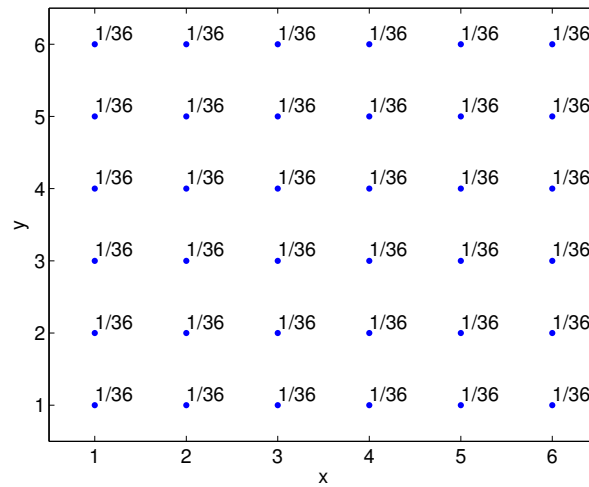


Figure 9.4: The probability mass function for rolling two dice.

Before we introduce key concepts that did not exist for a scalar random variable, let us give a simple example of joint probability distribution.

Example 9.2.1 rolling two dice

As the first example, let us consider rolling two dice. The first die takes on $x_i = i$, $i = 1, \dots, 6$, and the second die takes on $y_j = j$, $j = 1, \dots, 6$. The random vector associated with rolling the two dice is

$$(X, Y) ,$$

where X takes on $J_X = 6$ values and Y takes on $J_Y = 6$ values. Thus, the random vector (X, Y) takes on $J = J_X \cdot J_Y = 36$ values. Because (for a fair die) each of the 36 outcomes is equally likely, the probability mass function $f_{X,Y}$ is

$$f_{X,Y}(x_i, y_j) = \frac{1}{36}, \quad i = 1, \dots, 6, \quad j = 1, \dots, 6 .$$

The probability mass function is shown graphically in Figure 9.4.

9.2.2 Characterization of Joint Distributions

Now let us introduce few additional concepts useful for describing joint distributions. Throughout this section, we consider a random vector (X, Y) with the associated probability distribution $f_{X,Y}$. First is the *marginal density*, which is defined as

$$f_X(x_i) = \sum_{j=1}^{J_Y} f_{X,Y}(x_i, y_j), \quad i = 1, \dots, J_X .$$

In words, marginal density of X is the probability distribution of X disregarding Y . That is, we ignore the outcome of Y , and ask ourselves the question: How frequently does X take on the value x_i ? Clearly, this is equal to summing the joint probability $f_{X,Y}(x_i, y_j)$ for all values of y_j . Similarly, the marginal density for Y is

$$f_Y(y_j) = \sum_{i=1}^{J_X} f_{X,Y}(x_i, y_j), \quad j = 1, \dots, J_Y .$$

Again, in this case, we ignore the outcome of X and ask: How frequently does Y take on the value y_j ? Note that the marginal densities are valid probability distributions because

$$f_X(x_i) = \sum_{j=1}^{J_Y} f_{X,Y}(x_i, y_j) \leq \sum_{k=1}^{J_X} \sum_{j=1}^{J_Y} f_{X,Y}(x_k, y_j) = 1, \quad i = 1, \dots, J_X,$$

and

$$\sum_{i=1}^{J_X} f_X(x_i) = \sum_{i=1}^{J_X} \sum_{j=1}^{J_Y} f_{X,Y}(x_i, y_j) = 1.$$

The second concept is the *conditional probability*, which is the probability that X takes on the value x_i given Y has taken on the value y_j . The conditional probability is denoted by

$$f_{X|Y}(x_i|y_j), \quad i = 1, \dots, J_X, \quad \text{for a given } y_j.$$

The conditional probability can be expressed as

$$f_{X|Y}(x_i|y_j) = \frac{f_{X,Y}(x_i, y_j)}{f_Y(y_j)}.$$

In words, the probability that X takes on x_i given that Y has taken on y_j is equal to the probability that both events take on (x_i, y_j) normalized by the probability that Y takes on y_j disregarding x_i . We can consider a different interpretation of the relationship by rearranging the equation as

$$f_{X,Y}(x_i, y_j) = f_{X|Y}(x_i|y_j) f_Y(y_j) \tag{9.1}$$

and then summing on j to yield

$$f_X(x_i) = \sum_{j=1}^{J_Y} f_{X,Y}(x_i, y_j) = \sum_{j=1}^{J_Y} f_{X|Y}(x_i|y_j) f_Y(y_j).$$

In other words, the marginal probability of X taking on x_i is equal to the sum of the probabilities of X taking on x_i given Y has taken on y_j multiplied by the probability of Y taking on y_j disregarding x_i . We can also exchange the roles of x and y (which, combined with (9.1), yields “Bayes’ Law”).

Example 9.2.2 marginal and conditional density of rolling two dice

Let us revisit the example of rolling two dice, and illustrate how the marginal density and conditional density are computed. We recall that the probability mass function for the problem is

$$f_{X,Y}(x, y) = \frac{1}{36}, \quad x = 1, \dots, 6, \quad y = 1, \dots, 6.$$

The calculation of the marginal density of X is illustrated in Figure 9.5(a). For each x_i , $i = 1, \dots, 6$, we have

$$f_X(x_i) = \sum_{j=1}^6 f_{X,Y}(x_i, y_j) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6}, \quad i = 1, \dots, 6.$$

We can also deduce this from intuition and arrive at the same conclusion. Recall that marginal density of X is the probability density of X ignoring the outcome of Y . For this two-dice rolling



Figure 9.5: Illustration of calculating marginal density $f_X(x = 2)$ and $f_Y(y = 3)$.

example, it simply corresponds to the probability distribution of rolling a single die, which is clearly equal to

$$f_X(x) = \frac{1}{6}, \quad x = 1, \dots, 6 .$$

Similar calculation of the marginal density of Y , f_Y , is illustrated in Figure 9.5(b). In this case, ignoring the first die (X), the second die produces $y_j = j$, $j = 1, \dots, 6$, with the equal probability of $1/6$.

Let us now illustrate the calculation of conditional probability. As the first example, let us compute the conditional probability of X given Y . In particular, say we are given $y = 3$. As shown in Figure 9.6(a), the joint probability of everything except those corresponding to $y = 3$ are irrelevant (shaded region). Within the region with $y = 3$, we have six possible outcomes, each with the equal probability. Thus, we have

$$f_{X|Y}(x|y = 3) = \frac{1}{6}, \quad x = 1, \dots, 6 .$$

Note that, we can compute this by simply considering the select set of joint probabilities $f_{X,Y}(x, y = 3)$ and re-normalizing the probabilities by their sum. In other words,

$$f_{X|Y}(x|y = 3) = \frac{f_{X,Y}(x, y = 3)}{\sum_{i=1}^6 f_{X,Y}(x_i, y = 3)} = \frac{f_{X,Y}(x, y = 3)}{f_Y(y = 3)} ,$$

which is precisely equal to the formula we have introduced earlier.

Similarly, Figure 9.6(b) illustrates the calculation of the conditional probability $f_{Y|X}(y, x = 2)$. In this case, we only consider joint probability distribution of $f_{X,Y}(x = 2, y)$ and re-normalize the density by $f_X(x = 2)$.

A very important concept is *independence*. Two events are said to be independent if the occurrence of one event does not influence the outcome of the other event. More precisely, two random variables X and Y are said to be independent if their probability density function satisfies

$$f_{X,Y}(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j), \quad i = 1, \dots, J_X, \quad j = 1, \dots, J_Y .$$

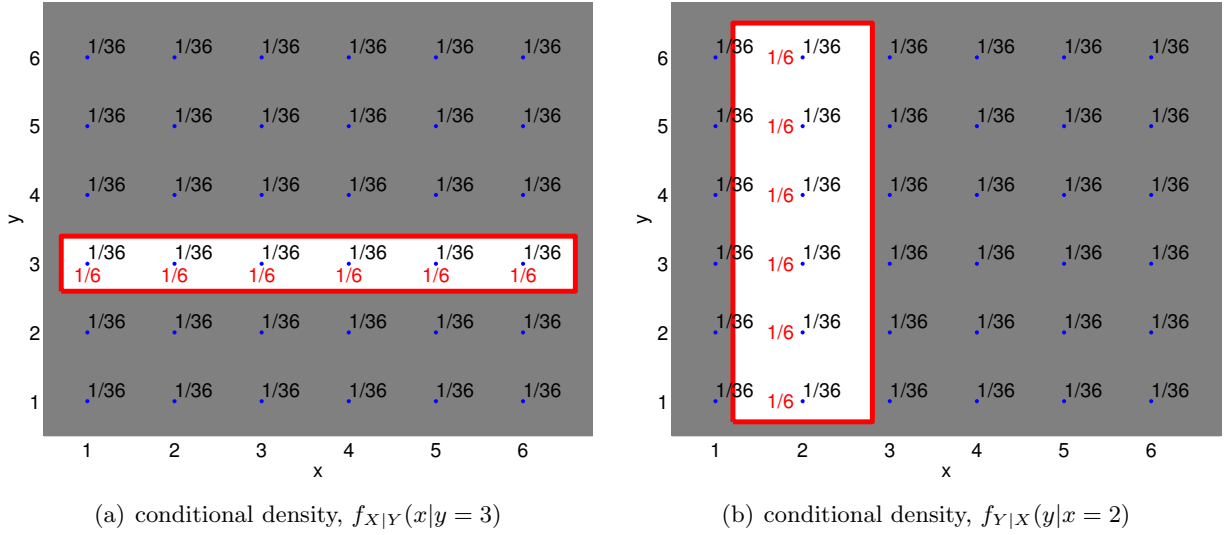


Figure 9.6: Illustration of calculating conditional density $f_{X|Y}(x|y=3)$ and $f_{Y|X}(y|x=2)$.

The fact that the probability density is simply a product of marginal densities means that we can draw X and Y separately according to their respective marginal probability and then form the random vector (X, Y) .

Using conditional probability, we can connect our intuitive understanding of independence with the precise definition. Namely,

$$f_{X|Y}(x_i|y_j) = \frac{f_{X,Y}(x_i, y_j)}{f_Y(y_j)} = \frac{f_X(x_i)f_Y(y_j)}{f_Y(y_j)} = f_X(x_i) .$$

That is, the conditional probability of X given Y is no different from the probability that X takes on x disregarding y . In other words, knowing the outcome of Y adds no additional information about the outcome of X . This agrees with our intuitive sense of independence.

To solidify the idea of independence, let us consider two canonical examples involving coin flips.

Example 9.2.3 independent events: two random variables associated with two independent coin flips

Let us consider flipping two fair coins. We associate the outcome of flipping the first and second coins with random variables X and Y , respectively. Furthermore, we associate the values of 1 and 0 to head and tail, respectively. We can associate the two flips with a random vector (X, Y) , whose possible outcomes are

$$(0, 0), \quad (0, 1), \quad (1, 0), \quad \text{and} \quad (1, 1) .$$

Intuitively, the two variables X and Y are independent because the outcome of the second flip, described by Y , is not influenced by the outcome of the first flip, described by X , and vice versa.

We can also show that X and Y are independent following the definition of independence. We are equally like to get any of the four outcomes, so the joint probability mass function is

$$f_{X,Y}(x, y) = \frac{1}{4}, \quad (x, y) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\} .$$

On the other hand, the marginal probability density of X is

$$f_X(x) = \frac{1}{2}, \quad x \in \{0, 1\}$$

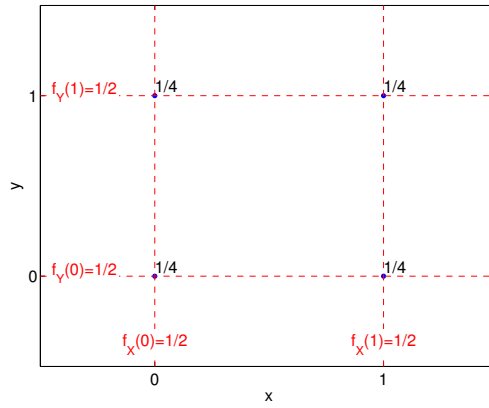


Figure 9.7: The probability mass function for flipping two independent coins.

because X is equally to produce a head or tail. Similarly, the marginal probability density of Y is

$$f_Y(y) = \frac{1}{2}, \quad y \in \{0, 1\} .$$

We note that

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) = \frac{1}{4}, \quad (x, y) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\} ,$$

which is the definition of independence.

The probability mass function of (X, Y) and the marginal density of X and Y are shown in Figure 9.7. The figure clearly shows that the joint density of (X, Y) is the product of the marginal density of X and Y . Let us show that this agrees with our intuition, in particular by considering the probability of $(X, Y) = (0, 0)$. First, the relative frequency that X takes on 0 is $1/2$. Second, of the events in which $x = 0$, $1/2$ of them take on $y = 0$. Note that this probability is independent of the value that X takes. Thus, the relative frequency of X taking on 0 and Y taking on 0 is $1/2$ of $1/2$, which is equal to $1/4$.

We can also consider conditional probability of an event that X takes on 1 given that Y takes on 0. The conditional probability is

$$f_{X|Y}(x = 1|y = 0) = \frac{f_{X,Y}(x = 1, y = 0)}{f_Y(y = 0)} = \frac{1/4}{1/2} = \frac{1}{2} .$$

This probability is equal to the marginal probability of $f_X(x = 1)$. This agrees with our intuition; given that two events are independent, we gain no additional information about the outcome of X from knowing the outcome of Y .

Example 9.2.4 non-independent events: two random variables associated with a single coin flip

Let us now consider flipping a single coin. We associate a Bernoulli random variables X and Y with

$$X = \begin{matrix} 1, & \text{head} \\ 0, & \text{tail} \end{matrix} \quad \text{and} \quad Y = \begin{matrix} 1, & \text{tail} \\ 0, & \text{head} \end{matrix} .$$

Note that a head results in $(X, Y) = (1, 0)$, whereas a tail results in $(X, Y) = (0, 1)$. Intuitively, the random variables are not independent, because the outcome of X completely determines Y , i.e. $X + Y = 1$.

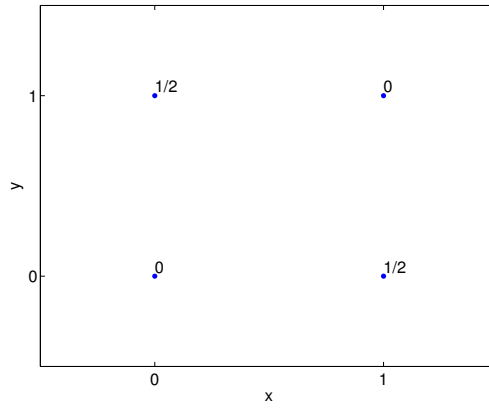


Figure 9.8: The probability mass function for flipping two independent coins.

Let us show that these two variables are not independent. We are equally like to get a head, $(1, 0)$, or a tail, $(0, 1)$. We cannot produce $(0, 0)$, because the coin cannot be head and tail at the same time. Similarly, $(1, 1)$ has probably of zero. Thus, the joint probability density function is

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{2}, & (x, y) = (0, 1) \\ \frac{1}{2}, & (x, y) = (1, 0) \\ 0, & (x, y) = (0, 0) \text{ or } (x, y) = (1, 1) . \end{cases}$$

The probability mass function is illustrated in Figure 9.8.

The marginal density of each of the event is the same as before, i.e. X is equally likely to take on 0 or 1, and Y is equally like to take on 0 or 1. Thus, we have

$$\begin{aligned} f_X(x) &= \frac{1}{2}, & x \in \{0, 1\} \\ f_Y(y) &= \frac{1}{2}, & y \in \{0, 1\} . \end{aligned}$$

For $(x, y) = (0, 0)$, we have

$$f_{X,Y}(x, y) = 0 \neq \frac{1}{4} = f_X(x) \cdot f_Y(y) .$$

So, X and Y are not independent.

We can also consider conditional probabilities. The conditional probability of $x = 1$ given that $y = 0$ is

$$f_{X|Y}(x = 1|y = 0) = \frac{f_{X,Y}(x = 1, y = 0)}{f_Y(y = 0)} = \frac{1/2}{1/2} = 1 .$$

In words, given that we know Y takes on 0, we know that X takes on 1. On the other hand, the conditional probability of $x = 1$ given that $y = 1$ is

$$f_{X|Y}(x = 0|y = 0) = \frac{f_{X,Y}(x = 0, y = 0)}{f_Y(y = 0)} = \frac{0}{1/2} = 0 .$$

In words, given that Y takes on 1, there is no way that X takes on 1. Unlike the previous example that associated (X, Y) with two independent coin flips, we know with certainty the outcome of X given the outcome of Y , and vice versa.

We have seen that independence is one way of describing the relationship between two events. Independence is a binary idea; either two events are independent or not independent. Another concept that describes how closely two events are related is correlation, which is a normalized covariance. The covariance of two random variables X and Y is denoted by $\text{Cov}(X, Y)$ and defined as

$$\text{Cov}(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)] .$$

The correlation of X and Y is denoted by ρ_{XY} and is defined as

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} ,$$

where we recall that σ_X and σ_Y are the standard deviation of X and Y , respectively. The correlation indicates how strongly two random events are related and takes on a value between -1 and 1 . In particular, two perfectly correlated events take on 1 (or -1), and two independent events take on 0 .

Two independent events have zero correlation because

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] = \sum_{j=1}^{J_Y} \sum_{i=1}^{J_X} (x_i - \mu_X)(y_j - \mu_Y) f_{X,Y}(x_i, y_j) \\ &= \sum_{j=1}^{J_Y} \sum_{i=1}^{J_X} (x_i - \mu_X)(y_j - \mu_Y) f_X(x_i) f_Y(y_j) \\ &= \left[\sum_{j=1}^{J_Y} (y_j - \mu_Y) f_Y(y_j) \right] \cdot \left[\sum_{i=1}^{J_X} (x_i - \mu_X) f_X(x_i) \right] \\ &= E[Y - \mu_Y] \cdot E[X - \mu_X] = 0 \cdot 0 = 0 . \end{aligned}$$

The third inequality follows from the definition of independence, $f_{X,Y}(x_i, y_j) = f_X(x_i) f_Y(y_j)$. Thus, if random variables are independent, then they are uncorrelated. However, the converse is not true in general.

9.3 Binomial Distribution

In the previous section, we saw random vectors consisting of two random variables, X and Y . Let us generalize the concept and introduce a random vector consisting of n components

$$(X_1, X_2, \dots, X_n) ,$$

where each X_i is a random variable. In particular, we are interested in the case where each X_i is a Bernoulli random variable with the probability of success of θ . Moreover, we assume that X_i , $i = 1, \dots, n$, are independent. Because the random variables are independent and each variable has the same distribution, they are said to be *independent and identically distributed* or i.i.d. for short.

Now, let us transform the i.i.d. random vector (X_1, \dots, X_n) of Bernoulli random variables to a random variable Z by summing its components, i.e.

$$Z_n = \sum_{i=1}^n X_i .$$

Note Z_n is simply a function of (X_1, X_2, \dots, X_n) , and in fact a simple function — the sum. Because X_i , $i = 1, \dots, n$, are random variables, their sum is also a random variable. In fact, this sum of Bernoulli random variable is called a *binomial random variable*. It is denoted by $Z_n \sim \mathcal{B}(n, \theta)$, where n is the number of Bernoulli events and θ is the probability of success of each event. Let us consider a few examples of binomial distributions.

Example 9.3.1 total number of heads in flipping two coins

Let us first revisit the case of flipping two coins. The random vector considered in this case is

$$(X_1, X_2) ,$$

where X_1 and X_2 are independent Bernoulli random variables associated with the first and second flip, respectively. As in the previous coin flip cases, we associate 1 with heads and 0 with tails. There are four possible outcome of these flips,

$$(0, 0), \quad (0, 1), \quad (1, 0), \quad \text{and} \quad (1, 1) .$$

From the two flips, we can construct the binomial distribution $Z_2 \sim \mathcal{B}(2, 1/2)$, corresponding to the total number of heads that results from flipping two coins. The binomial random variable is defined as

$$Z_2 = X_1 + X_2 .$$

Counting the number of heads associated with all possible outcomes of the coin flip, the binomial random variable takes on the following value:

First flip	0	0	1	1
Second flip	0	1	0	1
Z_2	0	1	1	2

Because the coin flips are independent and each coin flip has the probability density of $f_{X_i}(x) = 1/2$, $x = 0, 1$, their joint distribution is

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}, \quad (x_1, x_2) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\} .$$

In words, each of the four possible events are equally likely. Of these four equally likely events, $Z_2 \sim \mathcal{B}(2, 1/2)$ takes on the value of 0 in one event, 1 in two events, and 2 in one event. Thus, the behavior of the binomial random variable Z_2 can be concisely stated as

$$Z_2 = \begin{cases} 0, & \text{with probability } 1/4 \\ 1, & \text{with probability } 1/2 (= 2/4) \\ 2, & \text{with probability } 1/4 . \end{cases}$$

Note this example is very similar to Example 9.1.4: Z_2 , the sum of X_1 and X_2 , is our $g(X)$; we assign probabilities by invoking the mutually exclusive property, OR (union), and summation. Note that the mode, the value that Z_2 is most likely to take, is 1 for this case. The probability mass function of Z_2 is given by

$$f_{Z_2}(x) = \begin{cases} 1/4, & x = 0 \\ 1/2, & x = 1 \\ 1/4, & x = 2 . \end{cases}$$

Example 9.3.2 total number of heads in flipping three coins

Let us now extend the previous example to the case of flipping a fair coin three times. In this case, the random vector considered has three components,

$$(X_1, X_2, X_3) ,$$

with each X_i being a Bernoulli random variable with the probability of success of $1/2$. From the three flips, we can construct the binomial distribution $Z_3 \sim \mathcal{B}(3, 1/2)$ with

$$Z_3 = X_1 + X_2 + X_3 .$$

The all possible outcomes of the random vector and the associated outcomes of the binomial distribution are:

First flip	0	1	0	0	1	0	1	1
Second flip	0	0	1	0	1	1	0	1
Third flip	0	0	0	1	0	1	1	1
Z_3	0	1	1	1	2	2	2	3

Because the Bernoulli random variables are independent, their joint distribution is

$$f_{X_1, X_2, X_3}(x_1, x_2, x_3) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot f_{X_3}(x_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} .$$

In other words, each of the eight events is equally likely. Of the eight equally likely events, Z_3 takes on the value of 0 in one event, 1 in three events, 2 in three events, and 3 in one event. The behavior of the binomial variable Z_3 is summarized by

$$Z_3 = \begin{cases} 0, & \text{with probability } 1/8 \\ 1, & \text{with probability } 3/8 \\ 2, & \text{with probability } 3/8 \\ 3, & \text{with probability } 1/8 . \end{cases}$$

The probability mass function is thus given by

$$f_{Z_3}(x) = \begin{cases} 1/8, & x = 0 \\ 3/8, & x = 1 \\ 3/8, & x = 2 \\ 1/8, & x = 3 . \end{cases}$$

Example 9.3.3 total number of heads in flipping four coins

We can repeat the procedure for four flips of fair coins ($n = 4$ and $\theta = 1/2$). In this case, we consider the sum of the entries of a random vector consisting of four Bernoulli random variables, (X_1, X_2, X_3, X_4) . The behavior of $Z_4 = \mathcal{B}(4, 1/2)$ is summarized by

$$Z_4 = \begin{cases} 0, & \text{with probability } 1/16 \\ 1, & \text{with probability } 1/4 \\ 2, & \text{with probability } 3/8 \\ 3, & \text{with probability } 1/4 \\ 4, & \text{with probability } 1/16 . \end{cases}$$

Note that Z_4 is much more likely to take on the value of 2 than 0, because there are many equally-likely events that leads to $Z_4 = 2$, whereas there is only one event that leads to $Z_4 = 0$. In general, as the number of flips increase, the deviation of $Z_n \sim \mathcal{B}(n, \theta)$ from $n\theta$ becomes increasingly unlikely.

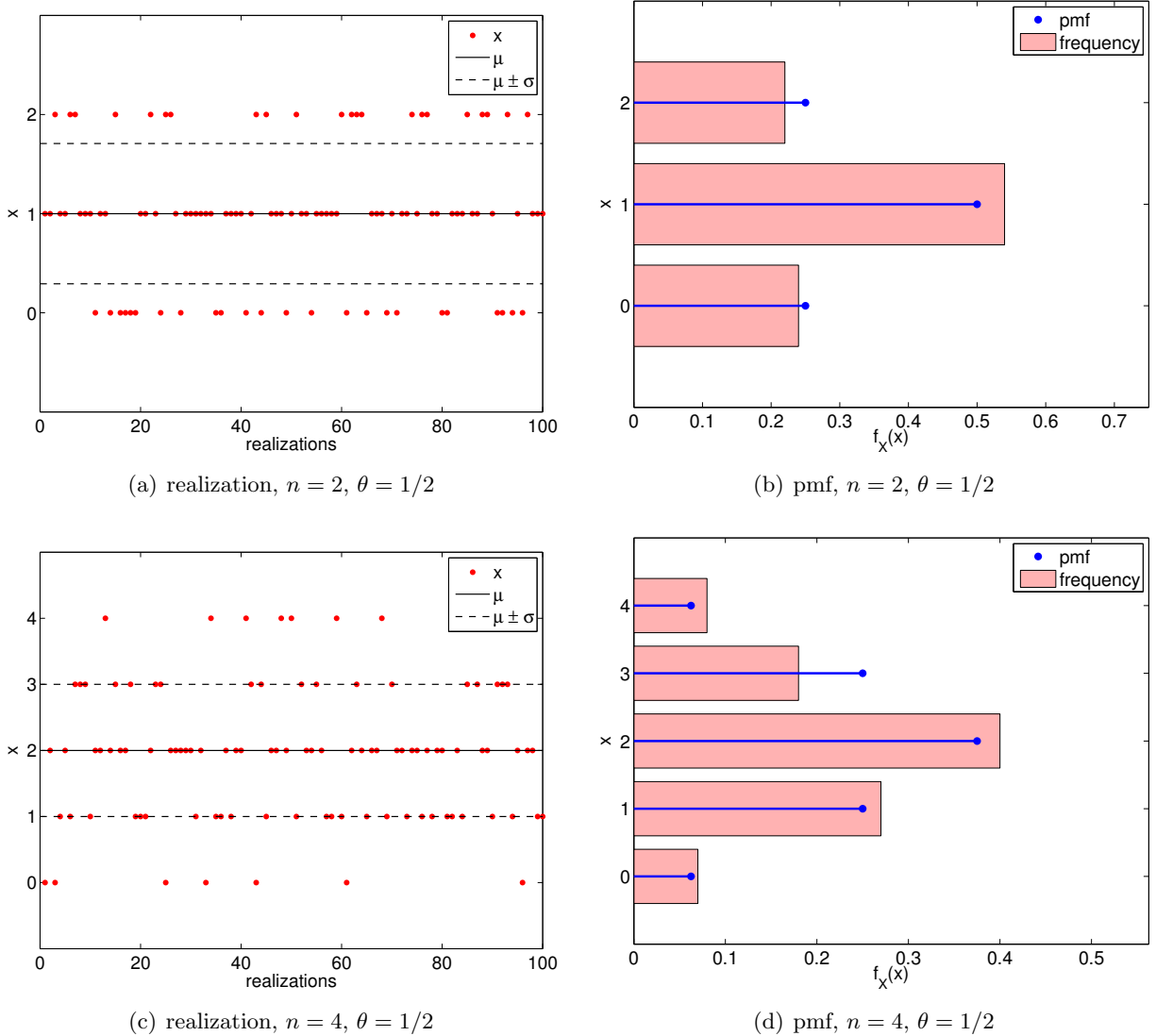


Figure 9.9: Illustration of the values taken by binomial random variables.

Figure 9.9 illustrates the values taken by binomial random variables for $n = 2$ and $n = 4$, both with $\theta = 1/2$. The histogram confirms that the distribution is more likely to take on the values near the mean because there are more sequences of the coin flips that realizes these values. We also note that the values become more concentrated near the mean, $n\theta$, relative to the range of the values it can take, $[0, n]$, as n increases. This is reflected in the decrease in the standard deviation relative to the width of the range of the values Z_n can take.

In general, binomial random variable $Z_n \sim \mathcal{B}(n, \theta)$ behaves as

$$Z_n = k, \quad \text{with probability} \quad \binom{n}{k} \theta^k (1 - \theta)^{n-k},$$

where $k = 1, \dots, n$. Recall that $\binom{n}{k}$ is the binomial coefficient, read “ n choose k ”: the number

of ways of picking k unordered outcomes from n possibilities. The value can be evaluated as

$$\frac{n}{k} \equiv \frac{n!}{(n-k)!k!} , \quad (9.2)$$

where $!$ denotes the factorial.

We can readily derive the formula for $\mathcal{B}(n, \theta)$. We think of n tosses as a binary number with n bits, and we ask how many ways k ones can appear. We can place the first one in n different places, the second one in $n-1$ different places, \dots , which yields $n!/(n-k)!$ possibilities. But since we are just counting the number of ones, the order of appearance does not matter, and we must divide $n!/(n-k)!$ by the number of different orders in which we can construct the same pattern of k ones — the first one can appear in k places, the second one in $k-1$ places, \dots , which yields $k!$. Thus there are “ n choose k ” ways to obtain k ones in a binary number of length n , or equivalently “ n choose k ” different binary numbers with k ones. Next, by independence, each pattern of k ones (and hence $n-k$ zeros) has probability $\theta^k(1-\theta)^{n-k}$. Finally, by the mutually exclusive property, the probability that $Z_n = k$ is simply the number of patterns with k ones multiplied by the probability that each such pattern occurs (note the probability is the same for each such pattern).

The mean and variance of a binomial distribution is given by

$$E[Z_n] = n\theta \quad \text{and} \quad \text{Var}[Z_n] = n\theta(1-\theta) .$$

Proof. The proof for the mean follows from the linearity of expectation, i.e.

$$E[Z_n] = E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n \theta = n\theta .$$

Note we can readily prove that the expectation of the sum is the sum of the expectations. We consider the n -dimensional sum over the joint mass function of the X_i weighted — per the definition of expectation — by the sum of the X_i , $i = 1, \dots, n$. Then, for any given X_i , we factorize the joint mass function: $n-1$ of the sums then return unity, while the last sum gives the expectation of X_i . The proof for variance relies on the pairwise independence of the random variables

$$\begin{aligned} \text{Var}[Z_n] &= E[(Z_n - E[Z_n])^2] = E\left[\left\{\left(\sum_{i=1}^n X_i\right) - n\theta\right\}^2\right] = E\left[\left\{\sum_{i=1}^n (X_i - \theta)\right\}^2\right] \\ &= E\left[\sum_{i=1}^n (X_i - \theta)^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (X_i - \theta)(X_j - \theta)\right] \\ &= \sum_{i=1}^n E[(X_i - \theta)^2] + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \cancel{E[(X_i - \theta)(X_j - \theta)]} \\ &= \sum_{i=1}^n \text{Var}[X_i] = \sum_{i=1}^n \theta(1-\theta) = n\theta(1-\theta) . \end{aligned}$$

The cross terms cancel because coin flips are independent. □

Note that the variance scales with n , or, equivalently, the standard deviation scales with \sqrt{n} . This turns out to be the key to Monte Carlo methods — numerical methods based on random variables — which is a focus of this unit.

Let us get some insight to how the general formula works by applying it to the binomial distribution associated with flipping coins three times.

Example 9.3.4 applying the general formula to coin flips

Let us revisit $Z_3 = \mathcal{B}(3, 1/2)$ associated with the number of heads in flipping three coins. The probability that $Z_3 = 0$ is, by substituting $n = 3$, $k = 0$, and $\theta = 1/2$,

$$f_{Z_3}(0) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} = \binom{3}{0} \left(\frac{1}{2}\right)^0 \left(1 - \frac{1}{2}\right)^{3-0} = \frac{3!}{0!(3-0)!} \left(\frac{1}{2}\right)^3 = \frac{1}{8},$$

which is consistent with the probability we obtained previously. Let us consider another case: the probability of $Z_3 = 2$ is

$$f_{Z_3}(2) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} = \binom{3}{2} \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{3-2} = \frac{3!}{2!(3-2)!} \left(\frac{1}{2}\right)^3 = \frac{3}{8}.$$

Note $3! = 6$ is the number of different ways in which we can form a binary number of length 3 with 2 ones if we care about order (in other words, the ones are distinguished by the order in which they are placed in the binary number): $(1_1, 1_2, 0)$, $(1_2, 1_1, 0)$, \dots ; and $2!$ is the number of patterns which are identical if we do not care about order (in other words, 1_1 and 1_2 can not be distinguished).

Note that the $\theta^k (1 - \theta)^{n-k}$ is the probability that the random vector of Bernoulli variables

$$(X_1, X_2, \dots, X_n),$$

realizes $X_1 = X_2 = \dots = X_k = 1$ and $X_{k+1} = \dots = X_n = 0$, and hence $Z_n = k$. Then, we multiply the probability with the number of different ways that we can realize the sum $Z_n = k$, which is equal to the number of different way of rearranging the random vector. Here, we are using the fact that the random variables are identically distributed. In the special case of (fair) coin flips, $\theta^k (1 - \theta)^{n-k} = (1/2)^k (1 - 1/2)^{n-k} = (1/2)^n$, because each random vector is equally likely.

9.4 Continuous Random Variables

9.4.1 Description

Let X be a random variable that takes on any real value in (say) an interval,

$$X \in [a, b].$$

The *probability density function* (pdf) is a function over $[a, b]$, $f_X(x)$, such that

$$f_X(x) \geq 0, \quad \forall x \in [a, b],$$

$$\int_a^b f_X(x) dx = 1.$$

Note that the condition that the probability mass function sums to unity is replaced by an integral condition for the continuous variable. The probability that X take on a value over an infinitesimal interval of length dx is

$$P(x \leq X \leq x + dx) = f_X(x) dx,$$

or, over a finite subinterval $[a', b'] \subset [a, b]$,

$$P(a' \leq X \leq b') = \int_{a'}^{b'} f_X(x) dx .$$

In other words, the probability that X takes on the value between a' and b' is the integral of the probability density function f_X over $[a', b']$. A particular instance of this is a *cumulative distribution function* (cdf), $F_X(x)$, which describes the probability that X will take on a value less than x , i.e.

$$F_X(x) = \int_a^x f_X(x) dx .$$

Note that any cdf satisfies the conditions

$$F_X(a) = \int_a^a f_X(x) dx = 0 \quad \text{and} \quad F_X(b) = \int_a^b f_X(x) dx = 1 .$$

Let us introduce a few notions useful for characterizing a pdf, and thus the behavior of the random variable. The *mean*, μ , or the expected value, $E[X]$, of the random variable X is

$$\mu = E[X] = \int_a^b f(x) x dx .$$

The *variance*, $\text{Var}(X)$, is a measure of the spread of the values that X takes about its mean and is defined by

$$\text{Var}(X) = E[(X - \mu)^2] = \int_a^b (x - \mu)^2 f(x) dx .$$

The variance can also be expressed as

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] = \int_a^b (x - \mu)^2 f(x) dx \\ &= \int_a^b x^2 f(x) dx - 2\mu \underbrace{\int_a^b x f(x) dx}_{\mu} + \mu^2 \int_a^b f(x) dx \\ &= E[X^2] - \mu^2 . \end{aligned}$$

Let us consider a few examples of continuous random variables.

Example 9.4.1 Uniform distribution

The uniform distribution is characterized by a constant pdf,

$$f_X(x; a, b) = \frac{1}{b - a} .$$

Note that the pdf satisfies the constraint

$$\int_a^b f_X(x; a, b) dx = \int_a^b \frac{1}{b - a} dx = 1 .$$

Furthermore, the probability that the random variable takes on a value in the subinterval $[a', b'] \in [a, b]$ is

$$P(a' \leq X \leq b') = \int_{a'}^{b'} f_X(x; a, b) dx = \int_{a'}^{b'} \frac{1}{b - a} dx = \frac{b' - a'}{b - a} .$$

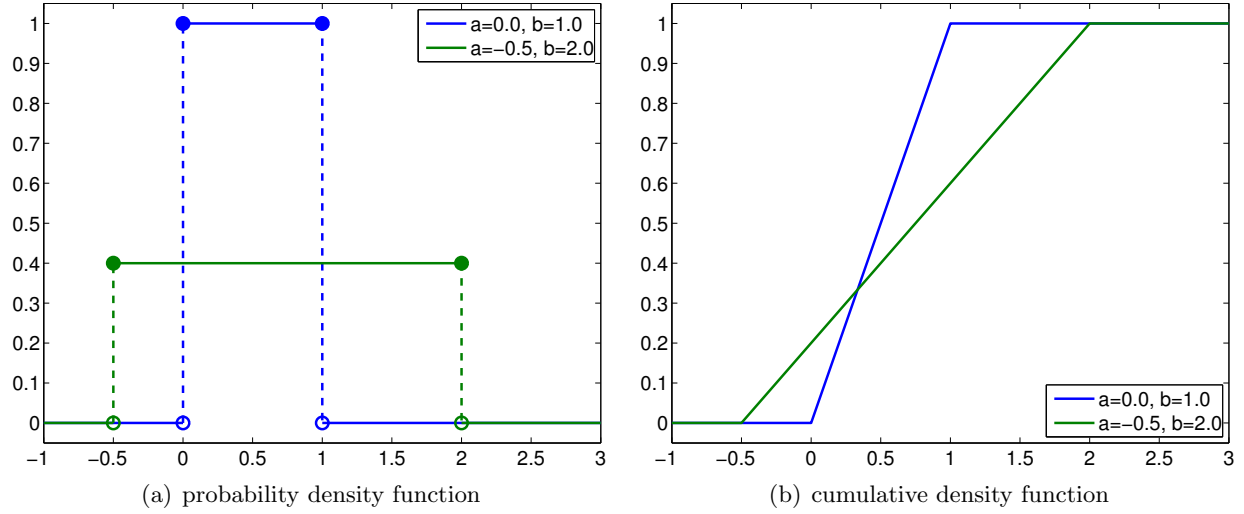


Figure 9.10: Uniform distributions

In other words, the probability that $X \in [a', b']$ is proportional to the relative length of the interval as the density is equally distributed. The distribution is compactly denoted as $\mathcal{U}(a, b)$ and we write $X \sim \mathcal{U}(a, b)$. The pdf and cdf for a few uniform distributions are shown in Figure 9.10.

An example of the values taken by a uniform random variable $\mathcal{U}(0, 1)$ is shown in Figure 9.11(a). By construction, the range of values that the variable takes is limited to between $a = 0$ and $b = 1$. As expected, there is no obvious concentration of the values within the range $[a, b]$. Figure 9.11(b) shows a histogram that summarizes the frequency of the event that X resides in bins $[x_i, x_i + \delta x]$, $i = 1, \dots, n_{\text{bin}}$. The relative frequency of occurrence is normalized by δx to be consistent with the definition of the probability density function. In particular, the integral of the region filled by the histogram is unity. While there is some spread in the frequencies of occurrences due to the relatively small sample size, the histogram resembles the probability density function. This is consistent with the frequentistic interpretation of probability.

The mean of the uniform distribution is given by

$$E[X] = \int_a^b x f_X(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{1}{2}(a+b) .$$

This agrees with our intuition, because if X is to take on a value between a and b with equal probability, then the mean would be the midpoint of the interval. The variance of the uniform distribution is

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 = \int_a^b x^2 f_X(x) dx - \left(\frac{1}{2}(a+b) \right)^2 \\ &= \int_a^b \frac{x^2}{b-a} dx - \frac{1}{2}(a+b)^2 = \frac{1}{12}(b-a)^2 . \end{aligned}$$

Example 9.4.2 Normal distribution

A normal distribution is characterized by the probability density function of the form

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left(-\frac{(x-\mu)^2}{2\sigma^2} \right) .$$

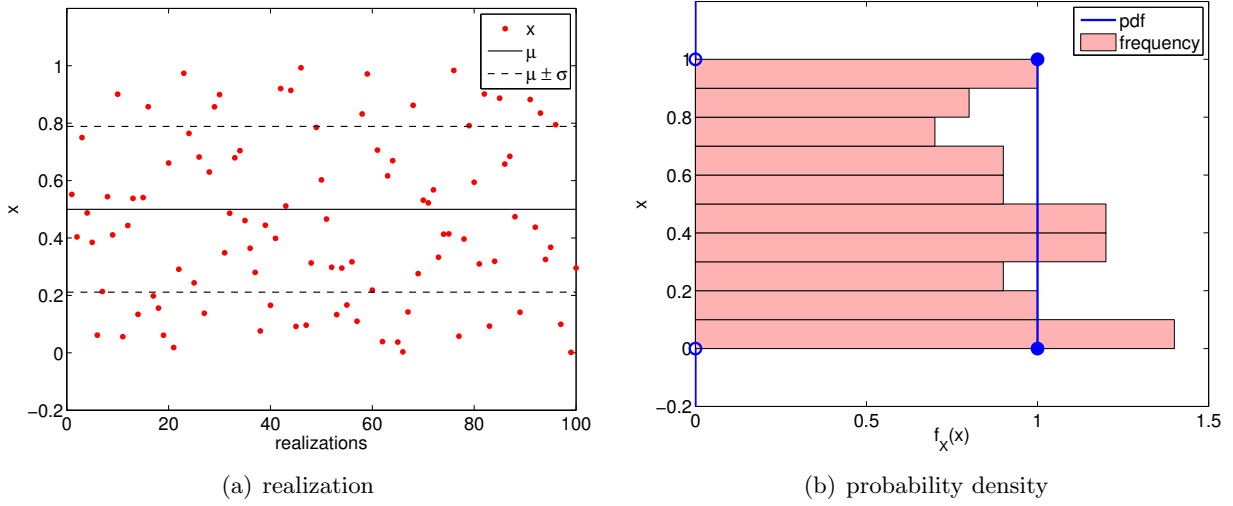


Figure 9.11: Illustration of the values taken by an uniform random variable ($a = 0$, $b = 1$).

The pdf is parameterized by two variables, the mean μ and the variance σ^2 . Note that the support of the distribution is the entire real axis, i.e. $a = -\infty$ and $b = \infty$. The normal distribution is concisely denoted by $\mathcal{N}(\mu, \sigma^2)$. The pdf and cdf for a few normal distributions are shown in Figure 9.12.

An example of the values taken by a normal random variable is shown in Figure 9.13. In principal, because the pdf has an infinite support, X can take on any real value; however, in practice, as the Gaussian function decays quickly away from the mean, the probability of X taking on a value many standard deviations away from the mean is small. Figure 9.13 clearly illustrates that the values taken by X is clustered near the mean. In particular, we can deduce from the cdf that X takes on the values within σ , 2σ , and 3σ with probability 68%, 95%, and 99.7%, respectively. In other words, the chance of X taking of the value outside of $\mu \pm 3\sigma$ is only 0.3%.

9.4.2 Transformations of Continuous Random Variables

Just like the discrete random variables, continuous random variables can be transformed by a function. The transformation of a random variable X by a function g produces another random variable, Y , and we denote this by

$$Y = g(X) .$$

We shall consider here only monotonic functions g .

Recall that we have described the random variable X by distribution

$$P(x \leq X \leq x + dx) = f_X(x) dx .$$

The transformed variable follows

$$P(y \leq Y \leq y + dy) = f_Y(y) dy .$$

Substitution of $y = g(x)$ and $dy = g'(x)dx$ and noting $g(x) + g'(x)dx = g(x + dx)$ results in

$$\begin{aligned} f_Y(y) dy &= P(g(x) \leq g(X) \leq g(x) + g'(x) dx) = P(g(x) \leq g(X) \leq g(x + dx)) \\ &= P(x \leq X \leq x + dx) = f_X(x) dx . \end{aligned}$$

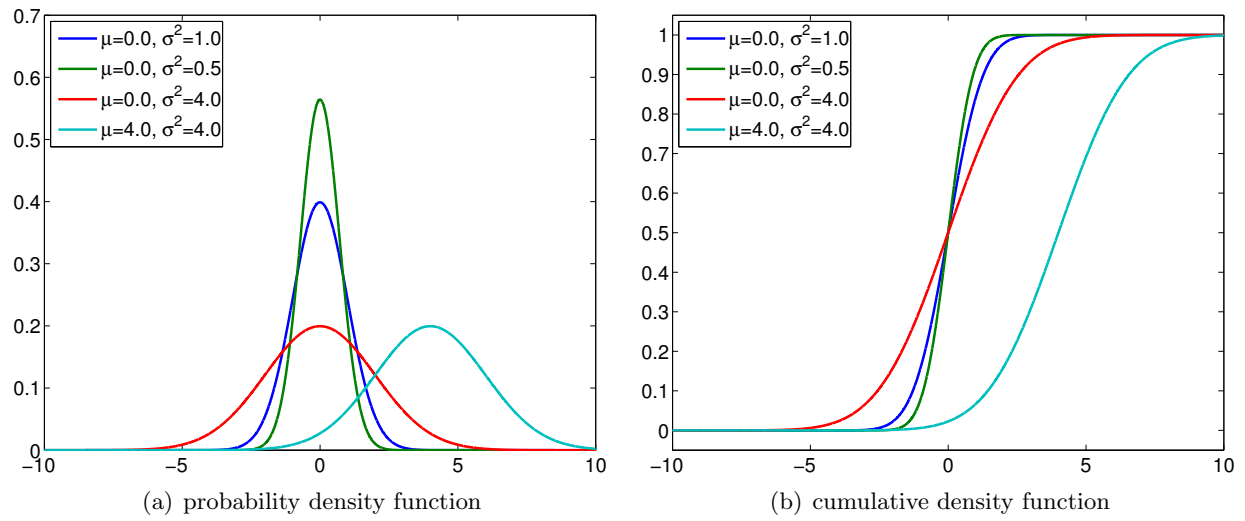


Figure 9.12: Normal distributions

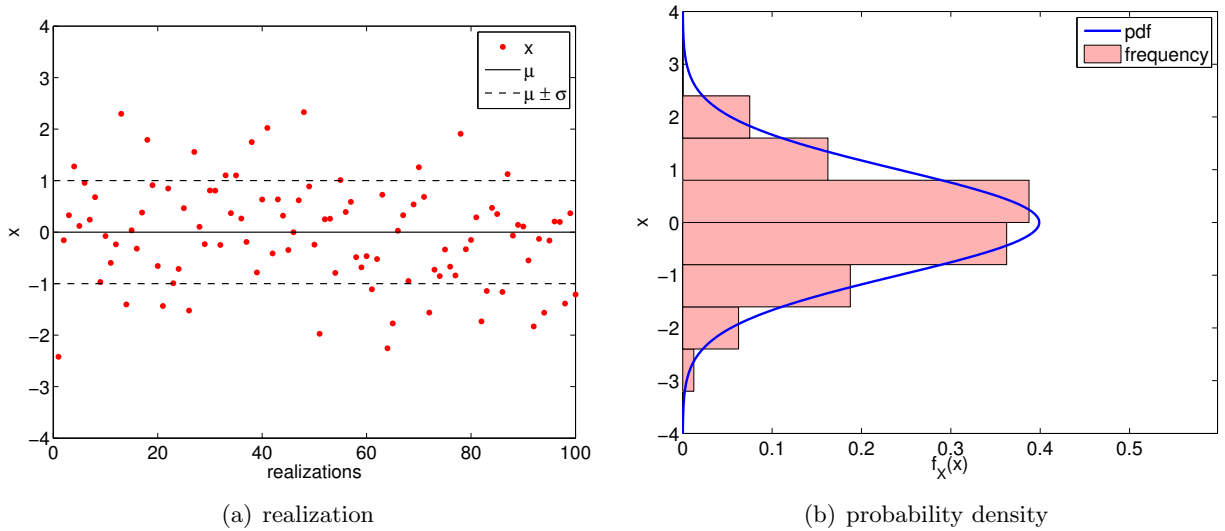


Figure 9.13: Illustration of the values taken by a normal random variable ($\mu = 0, \sigma = 1$).

In other words, $f_Y(y)dy = f_X(x) dx$. This is the continuous analog to $f_Y(y_j) = p_j = f_X(x_j)$ in the discrete case.

We can manipulate the expression to obtain an explicit expression for f_Y in terms of f_X and g . First we note that

$$y = g(x) \quad \Rightarrow \quad x = g^{-1}(y) \quad \text{and} \quad dx = \frac{dg^{-1}}{dy} dy .$$

Substitution of the expressions in $f_Y(y)dy = f_X(x) dx$ yields

$$f_Y(y) dy = f_X(x) dx = f_X(g^{-1}(y)) \cdot \frac{dg^{-1}}{dy} dy$$

or,

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{dg^{-1}}{dy} .$$

Conversely, we can also obtain an explicit expression for f_X in terms of f_Y and g . From $y = g(x)$ and $dy = g'(x) dx$, we obtain

$$f_X(x) dx = f_Y(y) dy = f_Y(g(x)) \cdot g'(x) dx \quad \Rightarrow \quad f_X(x) = f_Y(g(x)) \cdot g'(x) .$$

Assuming X takes on a value between a and b , and Y takes on a value between c and d , the mean of Y is

$$E[Y] = \int_c^d y f_Y(y) dy = \int_a^b g(x) f_X(x) dx ,$$

where the second equality follows from $f_Y(y) dy = f_X(x) dx$ and $y = g(x)$.

Example 9.4.3 Standard uniform distribution to a general uniform distribution

As the first example, let us consider the standard uniform distribution $U \sim \mathcal{U}(0, 1)$. We wish to generate a general uniform distribution $X \sim \mathcal{U}(a, b)$ defined on the interval $[a, b]$. Because a uniform distribution is unique determined by the two end points, we simply need to map the end point 0 to a and the point 1 to b . This is accomplished by the transformation

$$g(u) = a + (b - a)u .$$

Thus, $X \sim \mathcal{U}(a, b)$ is obtained by mapping $U \sim \mathcal{U}(0, 1)$ as

$$X = a + (b - a)U .$$

Proof. Proof follows directly from the transformation of the probability density function. The probability density function of U is

$$f_U(u) = \begin{cases} 1, & u \in [0, 1] \\ 0, & \text{otherwise} \end{cases} .$$

The inverse of the transformation $x = g(u) = a + (b - a)u$ is

$$g^{-1}(x) = \frac{x - a}{b - a} .$$

From the transformation of the probability density function, f_X is

$$f_X(x) = f_U(g^{-1}(x)) \cdot \frac{dg^{-1}}{dx} = f_U \frac{x-a}{b-a} \cdot \frac{1}{b-a} .$$

We note that f_U evaluates to 1 if

$$0 \leq \frac{x-a}{b-a} \leq 1 \quad \Rightarrow \quad a \leq x \leq b ,$$

and f_U evaluates to 0 otherwise. Thus, f_X simplifies to

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} ,$$

which is precisely the probability density function of $\mathcal{U}(a, b)$. □

Example 9.4.4 Standard uniform distribution to a discrete distribution

The uniform distribution can also be mapped to a discrete random variable. Consider a discrete random variable Y takes on three values ($J = 3$), with

$$y_1 = 0, \quad y_2 = 2, \quad \text{and} \quad y_3 = 3$$

with probability

$$f_Y(y) = \begin{cases} 1/2, & y_1 = 0 \\ 1/4, & y_2 = 2 \\ 1/4, & y_3 = 3 \end{cases} .$$

To generate Y , we can consider a discontinuous function g . To get the desired discrete probability distribution, we subdivide the interval $[0, 1]$ into three subintervals of appropriate lengths. In particular, to generate Y , we consider

$$g(x) = \begin{cases} 0, & x \in [0, 1/2) \\ 2, & x \in [1/2, 3/4) \\ 3, & x \in [3/4, 1] \end{cases} .$$

If we consider $Y = g(U)$, we have

$$Y = \begin{cases} 0, & U \in [0, 1/2) \\ 2, & U \in [1/2, 3/4) \\ 3, & U \in [3/4, 1] \end{cases} .$$

Because the probability that the standard uniform random variable takes on a value within a subinterval $[a', b']$ is equal to

$$P(a' \leq U \leq b') = \frac{b' - a'}{1 - 0} = b' - a' ,$$

the probability that Y takes on 0 is $1/2 - 0 = 1/2$, on 2 is $3/4 - 1/2 = 1/4$, and on 3 is $1 - 3/4 = 1/4$. This gives the desired probability distribution of Y .

Example 9.4.5 Standard normal distribution to a general normal distribution

Suppose we have the standard normal distribution $Z \sim \mathcal{N}(0, 1)$ and wish to map it to a general normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ with the mean μ and the variance σ . The transformation is given by

$$X = \mu + \sigma Z .$$

Conversely, we can map any normal distribution to the standard normal distribution by

$$Z = \frac{X - \mu}{\sigma} .$$

Proof. The probability density function of the standard normal distribution $Z \sim \mathcal{N}(0, 1)$ is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp -\frac{z^2}{2} .$$

Using the transformation of the probability density and the inverse mapping, $z(x) = (x - \mu)/\sigma$, we obtain

$$f_X(x) = f_Z(z(x)) \frac{dz}{dx} = f_Z \frac{x - \mu}{\sigma} \cdot \frac{1}{\sigma} .$$

Substitution of the probability density function f_Z yields

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \cdot \frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x - \mu)^2}{2\sigma^2} ,$$

which is exactly the probability density function of $\mathcal{N}(\mu, \sigma^2)$. □

9.4.3 The Central Limit Theorem

The ubiquitousness of the normal distribution stems from the central limit theorem. The central limits theorem states that the sum of a sufficiently larger number of i.i.d. random variables tends to normal distribution. In other words, if an experiment is repeated a larger number of times, the outcome on average approaches a normal distribution. Specifically, given i.i.d. random variables X_i , $i = 1, \dots, N$, each with the mean $E[X_i] = \mu$ and variance $\text{Var}[X_i] = \sigma^2$, their sum converges to

$$\sum_{i=1}^N X_i \rightarrow \mathcal{N}(\mu N, \sigma^2 N), \quad \text{as } N \rightarrow \infty .$$

To illustrate the theorem, let us consider the sum of uniform random variables $X_i \sim \mathcal{U}(-1/2, 1/2)$. The mean and variance of the random variable are $E[X_i] = 0$ and $\text{Var}[X_i] = 1/3$, respectively. By central limit theorem, we expect their sum, Z_N , to approach

$$Z_N \equiv \sum_{i=1}^N X_i \rightarrow \mathcal{N}(\mu N, \sigma^2 N) = \mathcal{N}(0, N/3) \quad \text{as } N \rightarrow \infty .$$

The pdf of the sum Z_i , $i = 1, 2, 3$, and the normal distribution $\mathcal{N}(0, N/3)|_{N=3} = \mathcal{N}(0, 1)$ are shown in Figure 9.14(a). Even though the original uniform distribution ($N = 1$) is far from normal and

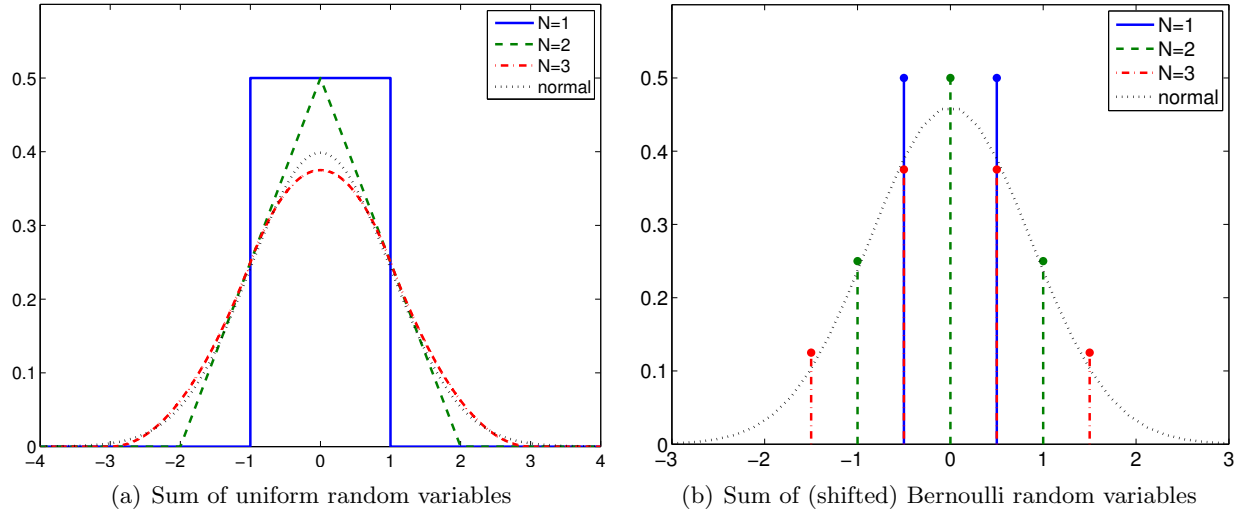


Figure 9.14: Illustration of the central limit theorem for continuous and discrete random variables.

$N = 3$ is not a large number, the pdf for $N = 3$ can be closely approximated by the normal distribution, confirming the central limit theorem.

The theorem also applies to discrete random variable. For example, let us consider the sum of (shifted) Bernoulli random variables,

$$X_i = \begin{cases} -1/2, & \text{with probability } 1/2 \\ 1/2, & \text{with probability } 1/2 \end{cases}.$$

Note that the value that X takes is shifted by $-1/2$ compared to the standard Bernoulli random variable, such that the variable has zero mean. The variance of the distribution is $\text{Var}[X_i] = 1/4$. As this is a discrete distribution, their sum also takes on discrete values; however, Figure 9.14(b) shows that the probability mass function can be closely approximated by the pdf for the normal distribution.

9.4.4 Generation of Pseudo-Random Numbers

To generate a realization of a random variable X computationally, we can use a pseudo-random number generator. Pseudo-random number generators are algorithms that generate a sequence of numbers that appear to be random. However, the actual sequence generated is completely determined by a seed — the variable that specifies the initial state of the generator. In other words, given a seed, the sequence of the numbers generated is completely deterministic and reproducible. Thus, to generate a different sequence each time, a pseudo-random number generator is seeded with something that is not fixed. A common choice is to use the current machine time. However, the deterministic nature of the pseudo-random number can be useful, for example, for debugging a code.

A typical computer language comes with a library that produces the standard continuous uniform distribution and the standard normal distribution. To generate other distributions, we can apply the transformations we considered earlier. For example, suppose that we have a pseudo-random number generator that generates the realization of $U \sim \mathcal{U}(0, 1)$,

$$u_1, u_2, \dots$$

Then, we can generate a realization of a general uniform distribution $X \sim \mathcal{U}(a, b)$,

$$x_1, x_2, \dots,$$

by using the transformation

$$x_i = a + (b - a)u_i, \quad i = 1, 2, \dots$$

Similarly, we can generate given a realization of the standard normal distribution $Z \sim \mathcal{N}(0, 1)$, z_1, z_2, \dots , we can generate a realization of a general normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, x_1, x_2, \dots , by

$$x_i = \mu + \sigma z_i, \quad i = 1, 2, \dots$$

Finally, if we wish to generate a discrete random number Y with the probability mass function

$$f_Y(y) = \begin{cases} 1/2, & y_1 = 0 \\ 1/4, & y_2 = 2 \\ 1/4, & y_3 = 3 \end{cases},$$

we can map a realization of the standard continuous uniform distribution $U \sim \mathcal{U}(0, 1)$, u_1, u_2, \dots , according to

$$y_i = \begin{cases} 0, & u_i \in [0, 1/2) \\ 2, & u_i \in [1/2, 3/4) \\ 3, & u_i \in [3/4, 1] \end{cases} \quad i = 1, 2, \dots$$

9.5 Continuous Random Vectors

Following the template used to extend discrete random variables to discrete random vectors, we now introduce the concept of continuous random vectors. Let $X = (X_1, X_2)$ be a random variable with

$$\begin{aligned} a_1 &\leq X_1 \leq b_1 \\ a_2 &\leq X_2 \leq b_2. \end{aligned}$$

The probability density function (pdf) is now a function over the rectangle

$$R \equiv [a_1, b_1] \times [a_2, b_2]$$

and is denoted by

$$f_{X_1, X_2}(x_1, x_2) \quad (\text{or, more concisely, } f_X(x_1, x_2)).$$

The pdf must satisfy the following conditions:

$$\begin{aligned} f_X(x_1, x_2) &\geq 0, \quad \forall (x_1, x_2) \in R \\ \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_X(x_1, x_2) &= 1. \end{aligned}$$

The value of the pdf can be interpreted as a probability per unit area, in the sense that

$$P(x_1 \leq X_1 \leq x_1 + dx_1, x_2 \leq X_2 \leq x_2 + dx_2) = f_X(x_1, x_2) dx_1 dx_2 ,$$

and

$$P(X \in D) = \int_D f_X(x_1, x_2) dx_1 dx_2 .$$

Let us now revisit key concepts used to characterize discrete joint distributions in the continuous setting. First, the *marginal density function* of X_1 is given by

$$f_{X_1}(x_1) = \int_{a_2}^{b_2} f_{X_1, X_2}(x_1, x_2) dx_2 .$$

Recall that the marginal density of X_1 describes the probability distribution of X_1 disregarding the state of X_2 . Similarly, the marginal density function of X_2 is

$$f_{X_2}(x_2) = \int_{a_1}^{b_1} f_{X_1, X_2}(x_1, x_2) dx_1 .$$

As in the discrete case, the marginal densities are also valid probability distributions.

The conditional probability density function of X_1 given X_2 is

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)} .$$

Similar to the discrete case, the marginal and conditional probabilities are related by

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1|X_2}(x_1|x_2) \cdot f_{X_2}(x_2) ,$$

or

$$f_{X_1}(x_1) = \int_{a_2}^{b_2} f_{X_1, X_2}(x_1, x_2) dx_2 = \int_{a_2}^{b_2} f_{X_1|X_2}(x_1|x_2) \cdot f_{X_2}(x_2) dx_2 .$$

In words, the marginal probability density function of X_1 is equal to the integration of the conditional probability density of f_{X_1, X_2} weighted by the probability density of X_2 .

Two continuous random variables are said to be independent if their joint probability density function satisfies

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) .$$

In terms of conditional probability, the independence means that

$$f_{X_1|X_2}(x_1, x_2) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{f_{X_1}(x_1) \cdot f_{X_2}(x_2)}{f_{X_2}(x_2)} = f_{X_1}(x_1) .$$

In words, knowing the outcome of X_2 does not add any new knowledge about the probability distribution of X_1 .

The covariance of X_1 and X_2 in the continuous case is defined as

$$\text{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] ,$$

and the correlation is given by

$$\rho_{X_1 X_2} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}} .$$

Recall that the correlation takes on a value between -1 and 1 and indicates how strongly the outcome of two random events are related. In particular, if the random variables are independent, then their correlation evaluates to zero. This is easily seen from

$$\begin{aligned} \text{Cov}(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)] = \int_{a_2}^{b_2} \int_{a_1}^{b_1} (x_1 - \mu_1)(x_2 - \mu_2) f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 \\ &= \int_{a_2}^{b_2} \int_{a_1}^{b_1} (x_1 - \mu_1)(x_2 - \mu_2) f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \\ &= \left[\int_{a_2}^{b_2} (x_2 - \mu_2) f_{X_2}(x_2) dx_2 \right] \cdot \left[\int_{a_1}^{b_1} (x_1 - \mu_1) f_{X_1}(x_1) dx_1 \right] \\ &= 0 \cdot 0 = 0 . \end{aligned}$$

Example 9.5.1 Bivariate uniform distribution

A bivariate uniform distribution is defined by two sets of parameters $[a_1, b_1]$ and $[a_2, b_2]$ that specify the range that X_1 and X_2 take on, respectively. The probability density function of (X_1, X_2) is

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{(b_1 - a_1)(b_2 - a_2)} .$$

Note here X_1 and X_2 are independent, so

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) ,$$

where

$$f_{X_1}(x_1) = \frac{1}{b_1 - a_1} \quad \text{and} \quad f_{X_2}(x_2) = \frac{1}{b_2 - a_2} .$$

As for the univariate case, we have

$$P(X \in D) = \frac{A_D}{A_R} ,$$

where A_D is the area of some arbitrary region D and A_R is the area of the rectangle. This relationship — together with our binomial distribution — will be the key ingredients for our Monte Carlo methods.

Note also that if A_D is itself a rectangle aligned with the coordinate directions, $A_D \equiv c_1 \leq x_1 \leq d_1, c_2 \leq x_2 \leq d_2$, then $P(X \in D)$ simplifies to the product of the length of D in x_1 , $(d_1 - c_1)$, divided by $b_1 - a_1$ and the length of D in x_2 , $(d_2 - c_2)$, divided by $b_2 - a_2$. Independence is manifested as a normalized product of lengths, or equivalently as the AND or intersection (not OR or union) of the two “event” rectangles $c_1 \leq x_1 \leq d_1, a_2 \leq x_2 \leq b_2$ and $a_1 \leq x_1 \leq b_1, c_2 \leq x_2 \leq d_2$.

To generate a realization of $X = (X_1, X_2)$, we express the vector as a function of two independent (scalar) uniform distributions. Namely, let us consider $U_1 \sim \mathcal{U}(0, 1)$ and $U_2 \sim \mathcal{U}(0, 1)$. Then, we can express the random vector as

$$\begin{aligned} X_1 &= a_1 + (b_1 - a_1)U_1 \\ X_2 &= a_2 + (b_2 - a_2)U_2 \\ X &= (X_1, X_2) . \end{aligned}$$

We stress that U_1 and U_2 must be independent in order for X_1 and X_2 to be independent.

Example 9.5.2 Bivariate normal distribution

A bivariate normal distribution is described by a probability density function of a form

$$f_{X_1, X_2}(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] \right\},$$

where (μ_1, μ_2) are the means, (σ_1^2, σ_2^2) are the variances, and ρ is the correlation. The pairs $\{\mu_1, \sigma_1^2\}$ and $\{\mu_2, \sigma_2^2\}$ describe the marginal distributions of X_1 and X_2 , respectively. The correlation coefficient must satisfy

$$-1 < \rho < 1$$

and, if $\rho = 0$, then X_1 and X_2 are uncorrelated. For a joint normal distribution, uncorrelated implies independence (this is not true for a general distribution).

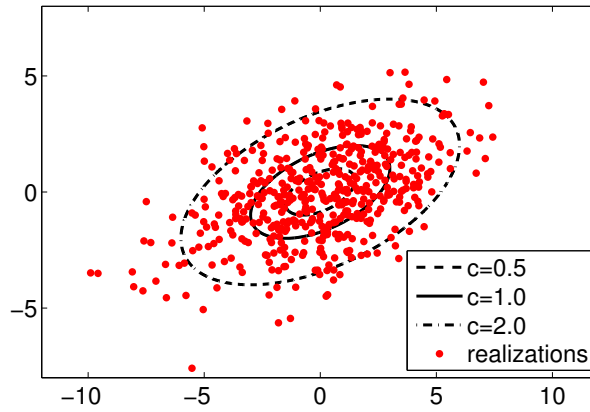


Figure 9.15: A bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 3$, $\sigma_2 = 2$, and $\rho = 1/2$.

The probability density function for the bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 3$, $\sigma_2 = 2$, and $\rho = 1/2$ is shown in Figure 9.15. The lines shown are the lines of equal density. In particular, the solid line corresponds to the 1σ line, and the dashed lines are for $\sigma/2$ and 2σ as indicated. 500 realizations of the distribution are also shown in red dots. For a bivariate distribution, the chances are 11.8%, 39.4%, and 86.5% that (X_1, X_2) takes on the value within $\sigma/2$, 1σ , and 2σ , respectively. The realizations shown confirm this trend, as only a small fraction of the red dots fall outside of the 2σ contour. This particular bivariate normal distribution has a weak positive correlation, i.e. given that X_2 is greater than its mean μ_{X_2} , there is a higher probability that X_1 is also greater than its mean, μ_{X_1} .

To understand the behavior of bivariate normal distributions in more detail, let us consider the marginal distributions of X_1 and X_2 . The marginal distribution of X_1 of a bivariate normal distribution characterized by $\{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\}$ is a univariate normal distribution with the mean μ_1 and the variance σ_1^2 , i.e.

$$f_{X_1}(x_1) \equiv \int_{x_2=-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_2 = f^{\text{normal}}(x_1; \mu_1, \sigma_1) .$$

In words, if we look at the samples of the binormal random variable (X_1, X_2) and focus on the behavior of X_1 only (i.e. disregard X_2), then we will observe that X_1 is normally distributed. Similarly, the marginal density of X_2 is

$$f_{X_2}(x_2) \equiv \int_{x_1=-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_1 = f^{\text{normal}}(x_2; \mu_2, \sigma_2) .$$

This rather surprising result is one of the properties of the binormal distribution, which in fact extends to higher-dimensional multivariate normal distributions.

Proof. For convenience, we will first rewrite the probability density function as

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \frac{1}{1-\rho^2} \exp \left[-\frac{1}{2} q(x_1, x_2) \right]$$

where the quadratic term is

$$q(x_1, x_2) = \frac{1}{1-\rho^2} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] .$$

We can manipulate the quadratic term to yield

$$\begin{aligned} q(x_1, x_2) &= \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{1}{1-\rho^2} \left[\frac{\rho^2(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] \\ &= \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{1}{1-\rho^2} \left[\frac{\rho(x_1 - \mu_1)}{\sigma_1} - \frac{x_2 - \mu_2}{\sigma_2} \right]^2 \\ &= \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{1}{\sigma_2^2(1-\rho^2)} \left[x_2 - \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1) \right]^2 . \end{aligned}$$

Substitution of the expression into the probability density function yields

$$\begin{aligned} f_{X_1, X_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \frac{1}{1-\rho^2} \exp \left[-\frac{1}{2} q(x_1, x_2) \right] \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left[-\frac{1}{2} \frac{(x_1 - \mu_1)^2}{\sigma_1^2} \right] \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma_2} \frac{1}{1-\rho^2} \exp \left[-\frac{1}{2} \frac{(x_2 - (\mu_2 + \rho(\sigma_2/\sigma_1)(x_1 - \mu_1)))^2}{\sigma_2^2(1-\rho^2)} \right] \\ &= f^{\text{normal}}(x_1; \mu_1, \sigma_1^2) \cdot f^{\text{normal}} \left(x_2; \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1), \sigma_2^2(1-\rho^2) \right) . \end{aligned}$$

Note that we have expressed the joint probability as the product of two univariate Gaussian functions. We caution that this does not imply independence, because the mean of the second distribution is dependent on the value of x_1 . Applying the definition of marginal density of X_1 and

integrating out the x_2 term, we obtain

$$\begin{aligned}
f_{X_1}(x_1) &= \int_{x_2=-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_2 \\
&= \int_{x_2=-\infty}^{\infty} f^{\text{normal}}(x_1; \mu_1, \sigma_1^2) \cdot f^{\text{normal}}\left(x_2; \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)\right) dx_2 \\
&= f^{\text{normal}}(x_1; \mu_1, \sigma_1^2) \cdot \int_{x_2=-\infty}^{\infty} f^{\text{normal}}\left(x_2; \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)\right) dx_2 \\
&= f^{\text{normal}}(x_1; \mu_1, \sigma_1^2) .
\end{aligned}$$

The integral of the second function evaluates to unity because it is a probability density function. Thus, the marginal density of X_1 is simply the univariate normal distribution with parameters μ_1 and σ_1 . The proof for the marginal density of X_2 is identical due to the symmetry of the joint probability density function. \square

Figure 9.16 shows the marginal densities f_{X_1} and f_{X_2} along with the $\sigma = 1$ - and $\sigma = 2$ -contours of the joint probability density. The dots superimposed on the joint density are 500 realizations of (X_1, X_2) . The histogram on the top summarizes the relative frequency of X_1 taking on a value within the bins for the 500 realizations. Similarly, the histogram on the right summarizes relative frequency of the values that X_2 takes. The histograms closely matches the theoretical marginal distributions for $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$. In particular, we note that the marginal densities are independent of the correlation coefficient ρ .

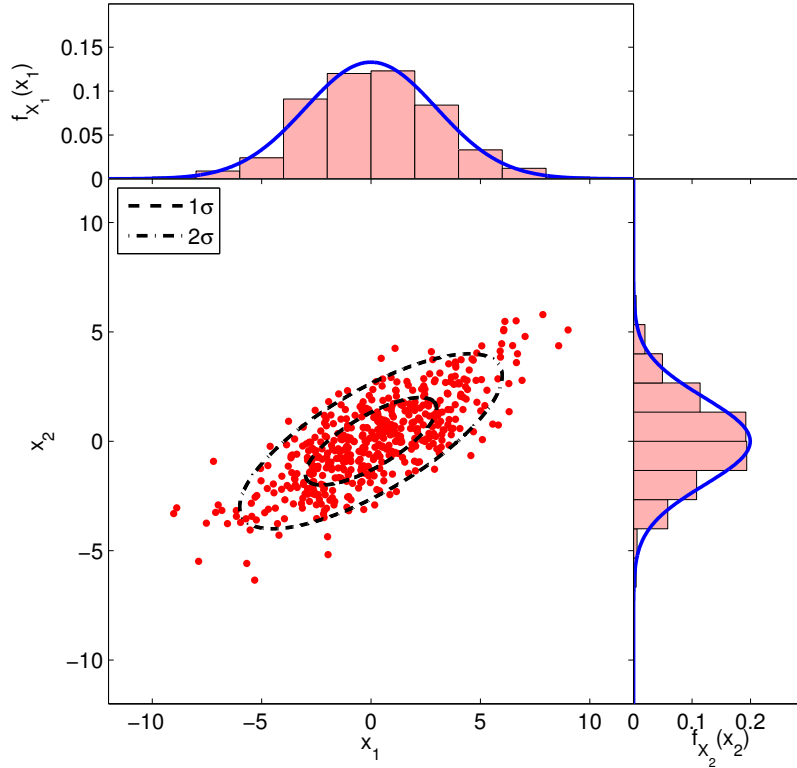


Figure 9.16: Illustration of marginal densities for a bivariate normal distribution ($\mu_1 = \mu_2 = 0$, $\sigma_1 = 3$, $\sigma_2 = 2$, $\rho = 3/4$).

Having studied the marginal densities of the bivariate normal distribution, let us now consider conditional probabilities. Combining the definition of conditional density and the expression for the joint and marginal densities, we obtain

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &= \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} = f^{\text{normal}}_{x_1; \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), (1 - \rho^2)\sigma_1^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \frac{1}{1 - \rho^2} \exp \left\{ -\frac{1}{2} \frac{(x_1 - (\mu_1 + \rho(\sigma_1/\sigma_2)x_2))^2}{\sigma_1^2(1 - \rho^2)} \right\}. \end{aligned}$$

Similarly, the conditional density of X_2 given X_1 is

$$f_{X_2|X_1}(x_2,x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} = f^{\text{normal}}_{x_2; \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), (1 - \rho^2)\sigma_2^2}.$$

Note that unlike the marginal probabilities, the conditional probabilities are function of the correlation coefficient ρ . In particular, the standard deviation of the conditional distribution (i.e. its spread about its mean) decreases with $|\rho|$ and vanishes as $\rho \rightarrow \pm 1$. In words, if the correlation is high, then we can deduce with a high probability the state of X_1 given the value that X_2 takes. We also note that the positive correlation ($\rho > 0$) results in the mean of the conditional probability $X_1|X_2$ shifted in the direction of X_2 . That is, if X_2 takes on a value higher than its mean, then it is more likely than not that X_1 takes on a value higher than its mean.

Proof. Starting with the definition of conditional probability and substituting the joint and marginal probability density functions,

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &= \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \frac{1}{1 - \rho^2} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] \right\} \\ &\quad \times \frac{\sqrt{2\pi}\sigma_2}{1} \exp \left\{ -\frac{1}{2} \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \frac{1}{1 - \rho^2} \exp \left\{ -\frac{1}{2} s(x_1, x_2) \right\} \end{aligned}$$

where

$$s(x_1, x_2) = \frac{1}{1 - \rho^2} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] - (1 - \rho^2) \frac{(x_2 - \mu_2)^2}{\sigma_2^2}.$$

Rearrangement of the quadratic term $s(x_1, x_2)$ yields

$$\begin{aligned} s(x_1, x_2) &= \frac{1}{1 - \rho^2} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{\rho^2(x_2 - \mu_2)^2}{\sigma_2^2} \right] \\ &= \frac{1}{1 - \rho^2} \left[\frac{x_1 - \mu_1}{\sigma_1} - \frac{\rho(x_2 - \mu_2)}{\sigma_2} \right]^2 \\ &= \frac{1}{\sigma_1^2(1 - \rho^2)} \left[x_1 - \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2) \right]^2. \end{aligned}$$

Substitution of the quadratic term into the conditional probability density function yields

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sqrt{2\pi}\sigma_1} \frac{1}{1-\rho^2} \exp \left\{ -\frac{1}{2} \frac{1}{\sigma_1^2(1-\rho^2)} \left(x_1 - \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2) \right)^2 \right\}$$

$$= f^{\text{normal}} \left(x_1; \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2), (1-\rho^2)\sigma_1^2 \right),$$

where the last equality follows from recognizing the univariate normal probability distribution function. \square

Figure 9.17 shows the conditional densities $f_{X_1|X_2}(x_1|x_2 = -2)$ and $f_{X_2|X_1}(x_2|x_1 = 3)$ for a bivariate normal distribution ($\mu_1 = \mu_2 = 0$, $\sigma_1 = 3$, $\sigma_2 = 2$, $\rho = 3/4$). The histograms are constructed by counting the relative frequency of occurrence for those realizations that falls near the conditional value of $x_2 = -2$ and $x_1 = 3$, respectively. Clearly, the mean of the conditional probability densities are shifted relative to the respective marginal densities. As $\rho = 3/4 > 0$ and $x_2 - \mu_2 = -2 < 0$, the mean for $X_1|X_2$ is shifted in the negative direction. Conversely, $\rho > 0$ and $x_1 - \mu_1 = 3 > 0$ shifts the mean for $X_2|X_1$ in the positive direction. We also note that the conditional probability densities are tighter than the respective marginal densities; due to the relative strong correlation of $\rho = 3/4$, we have a better knowledge of the one state when we know the value of the other state.

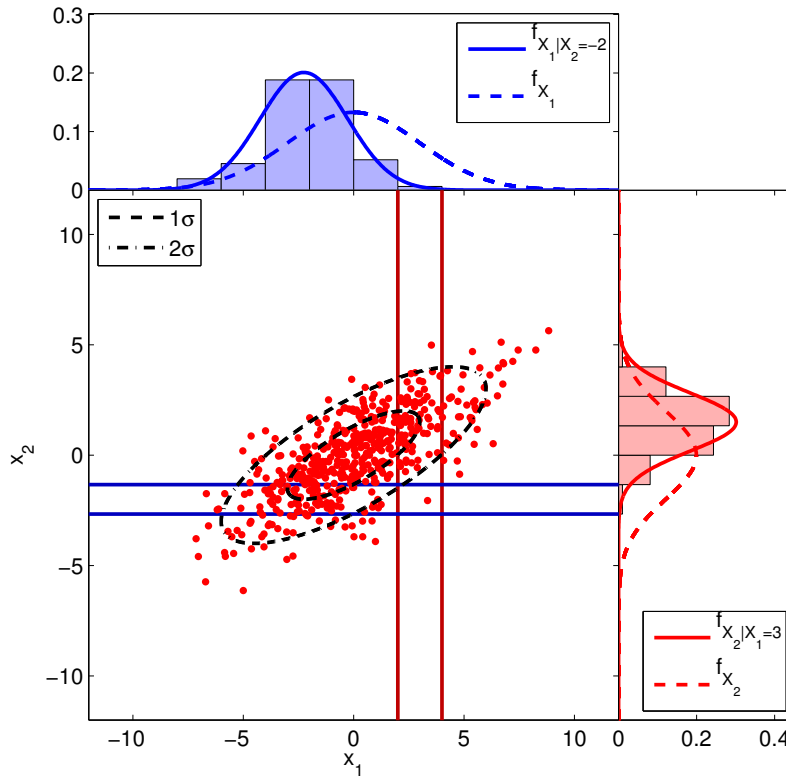


Figure 9.17: Illustration of conditional densities $f_{X_1|X_2}(x_1|x_2 = -2)$ and $f_{X_2|X_1}(x_2|x_1 = 3)$ for a bivariate normal distribution ($\mu_1 = \mu_2 = 0$, $\sigma_1 = 3$, $\sigma_2 = 2$, $\rho = 3/4$).

Finally, to solidify the idea of correlation, let us consider the 1σ -contour for bivariate normal distributions with several different values of ρ , shown in Figure 9.18. A stronger (positive) correlation implies that there is a high chance that a positive value of x_2 implies a positive value of

x_1 . Conversely, a strong negative correlation implies that there is a high chance a positive value of x_2 implies a negative value of x_1 . Zero correlation — which implies independence for normal distributions — means that we gain no additional information about the value that X_1 takes on by knowing the value of X_2 ; thus, the contour of equal probability density is not tilted.

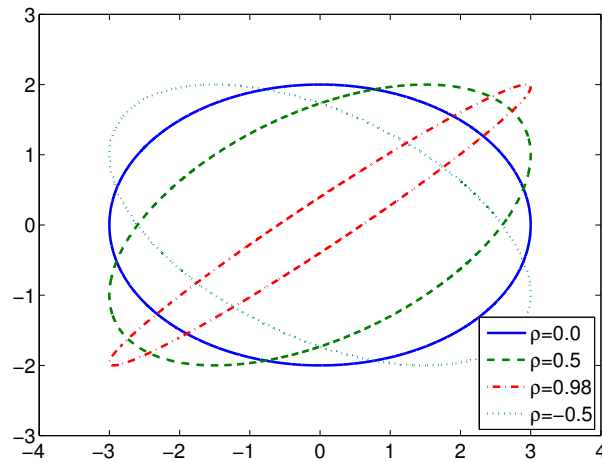


Figure 9.18: Bivariate normal distributions with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 3$, $\sigma_2 = 2$, and several values of ρ .

End Advanced Material

Chapter 10

Estimation: Bernoulli Coins

DRAFT V1.1 © The Authors. License: Creative Commons [BY-NC-SA 3.0](#)

10.1 The Sample Mean: An Estimator / Estimate

Let us illustrate the idea of sample mean in terms of a coin flip experiment, in which a coin is flipped n times. Unlike the previous cases, the coin may be unfair, i.e. the probability of head, θ , may not be equal to $1/2$. We assume that we do not know the value of θ , and we wish to estimate θ from data collected through n coin flips. In other words, this is a parameter estimation problem, where the unknown parameter is θ . Although this chapter serves as a prerequisite for subsequent chapters on Monte Carlo methods — in which we apply probabilistic concepts to calculate areas and more generally integrals — in fact the current chapter also illustrates how we might deduce physical parameters from noisy measurements. In short, statistics can be applied either to physical quantities treated as random variables or deterministic quantities which are re-interpreted as random (or pseudo-random).

As in the previous chapter, we associate the outcome of n flips with a random vector consisting of Bernoulli random variables,

$$(B_1, B_2, \dots, B_n) ,$$

where each B_i takes on the value of 1 with probability of θ and 0 with probability of $1 - \theta$. The random variables are i.i.d. because the outcome of one flip is independent of another flip and we are using the same coin.

We define the sample mean of n coin flips as

$$\bar{B}_n \equiv \frac{1}{n} \sum_{i=1}^n B_i ,$$

which is equal to the fraction of flips which are heads. Because \bar{B}_n is a transformation (i.e. sum) of random variables, it is also a random variable. Intuitively, given a large number of flips, we “expect” the fraction of flips which are heads — the frequency of heads — to approach the probability of a head, θ , for n sufficiently large. For this reason, the sample mean is our *estimator* in the context of parameter estimation. Because the estimator estimates the parameter θ , we will denote it by $\hat{\Theta}_n$, and it is given by

$$\hat{\Theta}_n = \bar{B}_n = \frac{1}{n} \sum_{i=1}^n B_i .$$

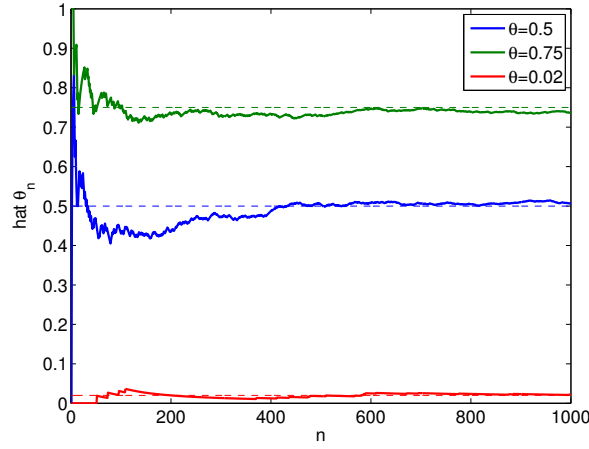


Figure 10.1: Convergence of estimate with n from a particular realization of coin flips.

We wish to estimate the parameter from a particular realization of coin flips (i.e. the random vector). For any particular realization, we calculate our estimate as

$$\hat{\theta}_n = \hat{b}_n \equiv \frac{1}{n} \sum_{i=1}^n b_i ,$$

where b_i is the particular outcome of the i -th flip. It is important to note that the b_i , $i = 1, \dots, n$, are numbers, each taking the value of either 0 or 1. Thus, $\hat{\theta}_n$ is a number and not a (random) distribution. Let us summarize the distinctions:

	r.v.?	Description
θ	no	Parameter to be estimated that governs the behavior of underlying distribution
$\hat{\Theta}_n$	yes	Estimator for the parameter θ
$\hat{\theta}_n$	no	Estimate for the parameter θ obtained from a particular realization

In general, how the random variable $\hat{\Theta}_n$ is distributed — in particular about θ — determines if $\hat{\Theta}_n$ is a good estimator for the parameter θ . An example of convergence of $\hat{\theta}_n$ to θ with n is shown in Figure 10.1. As n increases, $\hat{\theta}$ converges to θ for essentially all realization of B_i 's. This follows from the fact that $\hat{\Theta}_n$ is an unbiased estimator of θ — an estimator whose expected value is equal to the true parameter. We shall prove this shortly.

To gain a better insight into the behavior of $\hat{\Theta}_n$, we can construct the empirical distribution of $\hat{\Theta}_n$ by performing a large number of experiments for a given n . Let us denote the number of experiments by n_{exp} . In the first experiment, we work with a realization $(b_1, b_2, \dots, b_n)^{\text{exp } 1}$ and obtain the estimate by computing the mean, i.e.

$$\text{exp } 1 : (b_1, b_2, \dots, b_n)^{\text{exp } 1} \quad \Rightarrow \quad \bar{b}_n^{\text{exp } 1} = \frac{1}{n} \sum_{i=1}^n (b_i)^{\text{exp } 1} .$$

Similarly, for the second experiment, we work with a new realization to obtain

$$\text{exp } 2 : (b_1, b_2, \dots, b_n)^{\text{exp } 2} \quad \Rightarrow \quad \bar{b}_n^{\text{exp } 2} = \frac{1}{n} \sum_{i=1}^n (b_i)^{\text{exp } 2} .$$

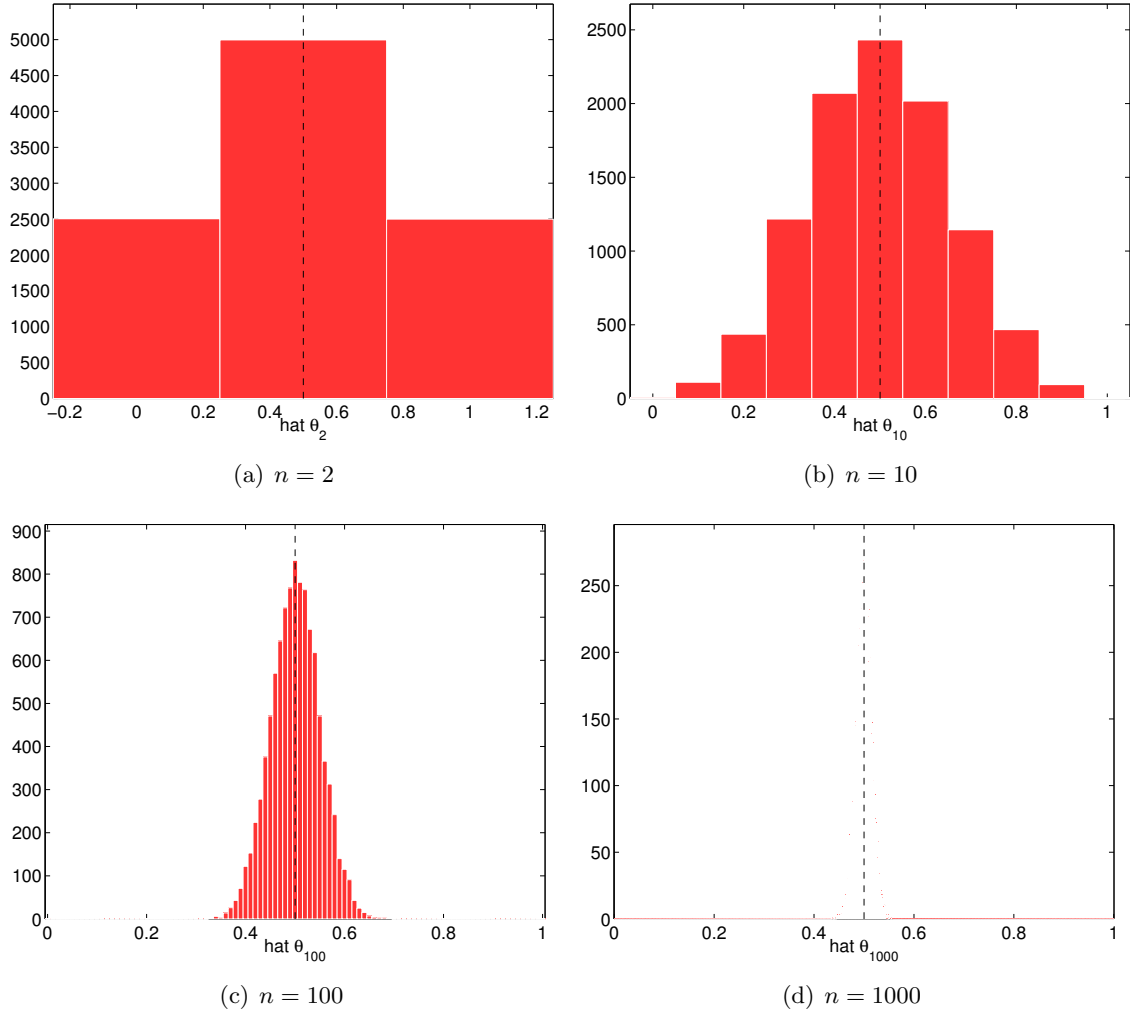


Figure 10.2: Empirical distribution of $\hat{\Theta}_n$ for $n = 2, 10, 100$, and 1000 and $\theta = 1/2$ obtained from 10,000 experiments.

Repeating the procedure n_{exp} times, we finally obtain

$$\text{exp } n_{\text{exp}} : (b_1, b_2, \dots, b_n)^{\text{exp } n_{\text{exp}}} \Rightarrow \bar{b}_n^{\text{exp } n_{\text{exp}}} = \frac{1}{n} \sum_{i=1}^n (b_i)^{\text{exp } n_{\text{exp}}} .$$

We note that \bar{b}_n can take any value k/n , $k = 0, \dots, n$. We can compute the frequency of \bar{b}_n taking on a certain value, i.e. the number of experiments that produces $\bar{b}_n = k/n$.

The numerical result of performing 10,000 experiments for $n = 2, 10, 100$, and 1000 flips are shown in Figure 10.2. The empirical distribution of $\hat{\Theta}_n$ shows that $\hat{\Theta}_n$ more frequently takes on the values close to the underlying parameter θ as the number of flips, n , increases. Thus, the numerical experiment confirms that $\hat{\Theta}_n$ is indeed a good estimator of θ if n is sufficiently large.

Having seen that our estimate converges to the true parameter θ in practice, we will now analyze the convergence behavior to the true parameter by relating the sample mean to a binomial distribution. Recall, that the binomial distribution represents the number of heads obtained in

flipping a coin n times, i.e. if $Z_n \sim \mathcal{B}(n, \theta)$, then

$$Z_n = \sum_{i=1}^n B_i ,$$

where B_i , $i = 1, \dots, n$, are the i.i.d. Bernoulli random variable representing the outcome of coin flips (each having the probability of head of θ). The binomial distribution and the sample mean are related by

$$\hat{\Theta}_n = \frac{1}{n} Z_n .$$

The mean (a deterministic parameter) of the sample mean (a random variable) is

$$E[\hat{\Theta}_n] = E\left[\frac{1}{n} Z_n\right] = \frac{1}{n} E[Z_n] = \frac{1}{n} (n\theta) = \theta .$$

In other words, $\hat{\Theta}_n$ is an unbiased estimator of θ . The variance of the sample mean is

$$\begin{aligned} \text{Var}[\hat{\Theta}_n] &= E[(\hat{\Theta}_n - E[\hat{\Theta}_n])^2] = E\left[\frac{1}{n} Z_n - \frac{1}{n} E[Z_n]\right]^2 = \frac{1}{n^2} E[(Z_n - E[Z_n])^2] \\ &= \frac{1}{n^2} \text{Var}[Z_n] = \frac{1}{n^2} n\theta(1 - \theta) = \frac{\theta(1 - \theta)}{n} . \end{aligned}$$

The standard deviation of $\hat{\Theta}_n$ is

$$\sigma_{\hat{\Theta}_n} = \sqrt{\text{Var}[\hat{\Theta}_n]} = \frac{\sqrt{\theta(1 - \theta)}}{n} .$$

Thus, the standard deviation of $\hat{\Theta}_n$ decreases with n , and in particular tends to zero as $1/\sqrt{n}$. This implies that $\hat{\Theta}_n \rightarrow \theta$ as $n \rightarrow \infty$ because it is very unlikely that $\hat{\Theta}_n$ will take on a value many standard deviations away from the mean. In other words, the estimator converges to the true parameter with the number of flips.

10.2 Confidence Intervals

Let us now introduce the concept of confidence interval. The confidence interval is a probabilistic *a posteriori* error bounds. *A posteriori* error bounds, as oppose to *a priori* error bounds, incorporate the information gathered in the experiment.

To understand the behavior of the estimator $\hat{\Theta}_n$, which is a random variable defined by

$$B_1, \dots, B_n \quad \Rightarrow \quad \hat{\Theta}_n = \bar{B}_n = \frac{1}{n} \sum_{i=1}^n B_i ,$$

we typically perform a single experiment in practice to generate a realization (b_1, \dots, b_n) . Then, we estimate the parameter by a number $\hat{\theta}_n$ given by

$$b_1, \dots, b_n \quad \Rightarrow \quad \hat{\theta}_n = \bar{b}_n = \frac{1}{n} \sum_{i=1}^n b_i .$$

A natural question is: how good is the estimate $\hat{\theta}_n$? How can we quantify the small deviations of $\hat{\theta}_n$ from θ as n increases? To answer these questions, we construct a confidence interval, $[\text{CI}]$, defined by

$$[\text{CI}]_n \equiv \left[\hat{\theta}_n - z \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}}, \hat{\theta}_n + z \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}} \right]$$

such that

$$P(\theta \in [\text{CI}]_n) = \gamma(z) .$$

We recall that θ is the true parameter; thus, γ is the confidence level that the true parameter falls within the confidence interval. Note that $[\text{CI}]_n$ is a random variable because $\hat{\theta}_n$ is a random variable.

As $n \rightarrow \infty$, $\hat{\theta}_n$ approaches a normal distribution due to the central limit theorem. Thus, for a large enough n , a typical confidence value of $\gamma = 0.95$ results in $z \approx 1.96$. In other words, if we use $z = 1.96$ to construct our confidence interval, there is a 95% probability that the true parameter lies within the confidence interval (for a sufficiently large n such that normality is justified). In general, as γ increases, z increases: if we want to ensure that the parameter lies within a confidence interval at a higher level of confidence, then the width of the confidence interval must be increased for a given n . The appearance of $1/\sqrt{n}$ in the confidence interval is due to the appearance of the $1/\sqrt{n}$ in the standard deviation of the estimator, $\sigma_{\hat{\theta}_n}$.

10.2.1 Frequentistic Interpretation

To get a better insight into the behavior of the confidence interval, let us provide an frequentistic interpolation of the interval¹. Let us perform n_{exp} experiments and construct n_{exp} realizations of confidence intervals, i.e.

$$[\text{ci}]_n^j = \left[\hat{\theta}_n^j - z \sqrt{\frac{\hat{\theta}_n^j(1 - \hat{\theta}_n^j)}{n}}, \hat{\theta}_n^j + z \sqrt{\frac{\hat{\theta}_n^j(1 - \hat{\theta}_n^j)}{n}} \right], \quad j = 1, \dots, n_{\text{exp}},$$

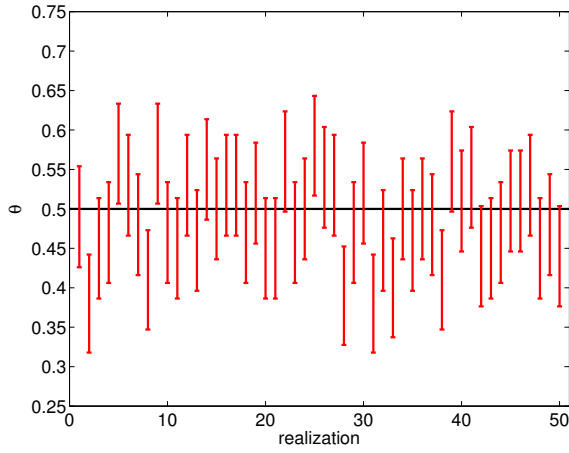
where the realization of sample means is given by

$$(b_1, \dots, b_n)^j \Rightarrow \hat{\theta}^j = \frac{1}{n} \sum_{i=1}^n b_i^j .$$

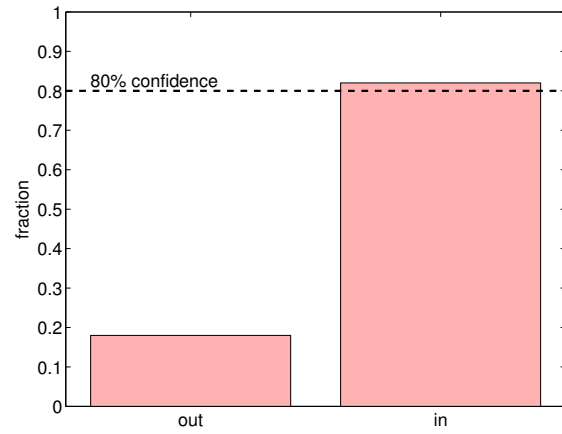
Then, as $n_{\text{exp}} \rightarrow \infty$, the fraction of experiments for which the true parameter θ lies inside $[\text{ci}]_n^j$ tends to γ .

An example of confidence intervals for estimating the mean of Bernoulli random variable ($\theta = 0.5$) using $n = 100$ samples is shown in Figure 10.3. In particular, we consider sets of 50 different realizations and construct 80% ($z = 1.28$) and 95% ($z = 1.96$) confidence intervals for each of the realizations. The histograms shown in Figure 10.3(b) and 10.3(d) summarize the relative frequency of the true parameter falling in and out of the confidence intervals. We observe that 80% and 95% confidence intervals include the true parameter θ in 82% (9/51) and 94% (47/50) of the realizations, respectively; the numbers are in good agreement with the frequentistic interpretation of the confidence intervals. Note that, for the same number of samples n , the 95% confidence interval has a larger width, as it must ensure that the true parameter lies within the interval with a higher probability.

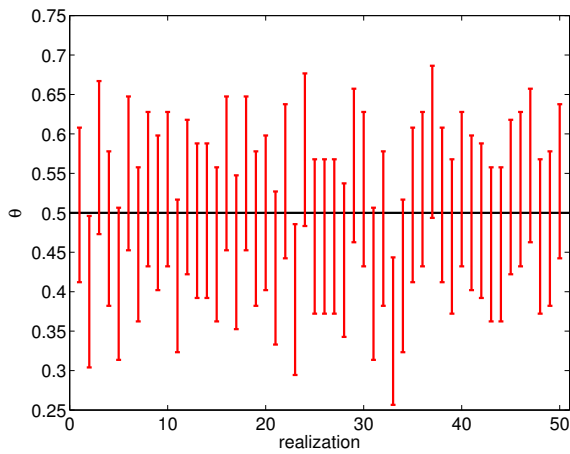
¹The other interpretation is based on the Bayesian view. We will not discuss the distinctions between the frequentist and Bayesian views.



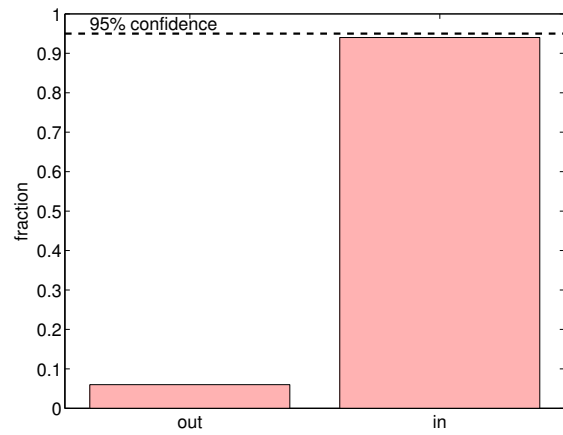
(a) 80% confidence



(b) 80% confidence in/out



(c) 95% confidence



(d) 95% confidence in/out

Figure 10.3: An example of confidence intervals for estimating the mean of a Bernoulli random variable ($\theta = 0.5$) using 100 samples.

10.2.2 Convergence

Let us define a relative error estimate in prediction for the true parameter θ as

$$\text{RelErr} \equiv \frac{\text{half length of [ci]}}{\hat{\theta}} = z \frac{\overline{1 - \hat{\theta}_n}}{\hat{\theta}_n n} .$$

The appearance of $1/\sqrt{n}$ convergence of the relative error is due to the $1/\sqrt{n}$ variation of the standard deviation $\sigma_{\hat{\theta}_n}$. Thus, the relative error converges in the sense that

$$\text{RelErr} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty .$$

However, the convergence rate is slow

$$\text{RelErr} \sim n^{-1/2} ,$$

i.e., the convergence rate is of order $1/2$ as $n \rightarrow \infty$. Moreover, note that rare events (i.e. low θ) are difficult to estimate accurately, as

$$\text{RelErr} \sim \hat{\theta}_n^{-1/2} .$$

This means that, if the number of experiments is fixed, the relative error in predicting an event that happens at 0.1% probability ($\theta = 0.001$) is 10 times larger than that for an event that happens at 10% probability ($\theta = 0.1$). Combined with the convergence rate of $n^{-1/2}$, it takes 100 times more experiments to achieve the similar level of relative error if the event is 100 times less likely. Thus, predicting the probability of a rare event is costly.

10.3 Cumulative Sample Means

In this subsection, we present a practical means of computing sample means. Let us denote the total number of coin flips by n_{\max} , which defines the size of our sample. We assume $n_{\text{exp}} = 1$, as is almost always the case in practice. We create our sample of size n_{\max} , and then for $n = 1, \dots, n_{\max}$ we compute a sequence of cumulative sample means. That is, we start with a realization of n_{\max} coin tosses,

$$b_1, b_2, \dots, b_n, \dots, b_{n_{\max}} ,$$

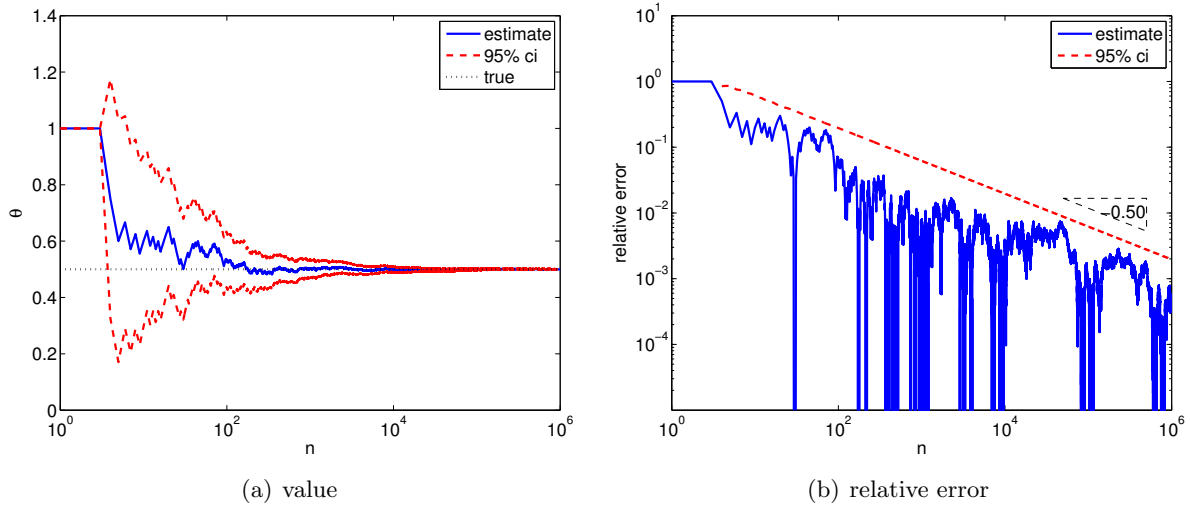


Figure 10.4: Cumulative sample mean, confidence intervals, and their convergence for a Bernoulli random variable ($\theta = 0.5$).

and then compute the cumulative values,

$$\begin{aligned}
 \hat{\theta}_1 &= \bar{b}_1 = \frac{1}{1} \cdot b_1 \quad \text{and} \quad [\text{ci}]_1 = \left[\hat{\theta}_1 - z \frac{\overline{\hat{\theta}_1(1 - \hat{\theta}_1)}}{1}, \hat{\theta}_1 + z \frac{\overline{\hat{\theta}_1(1 - \hat{\theta}_1)}}{1} \right] \\
 \hat{\theta}_2 &= \bar{b}_2 = \frac{1}{2}(b_1 + b_2) \quad \text{and} \quad [\text{ci}]_2 = \left[\hat{\theta}_2 - z \frac{\overline{\hat{\theta}_2(1 - \hat{\theta}_2)}}{2}, \hat{\theta}_2 + z \frac{\overline{\hat{\theta}_2(1 - \hat{\theta}_2)}}{2} \right] \\
 &\vdots \\
 \hat{\theta}_n &= \bar{b}_n = \frac{1}{n} \sum_{i=1}^n b_i \quad \text{and} \quad [\text{ci}]_n = \left[\hat{\theta}_n - z \frac{\overline{\hat{\theta}_n(1 - \hat{\theta}_n)}}{n}, \hat{\theta}_n + z \frac{\overline{\hat{\theta}_n(1 - \hat{\theta}_n)}}{n} \right] \\
 &\vdots \\
 \hat{\theta}_{n_{\max}} &= \bar{b}_{n_{\max}} = \frac{1}{n} \sum_{i=1}^{n_{\max}} b_i \quad \text{and} \quad [\text{ci}]_{n_{\max}} = \left[\hat{\theta}_{n_{\max}} - z \frac{\overline{\hat{\theta}_{n_{\max}}(1 - \hat{\theta}_{n_{\max}})}}{n_{\max}}, \hat{\theta}_{n_{\max}} + z \frac{\overline{\hat{\theta}_{n_{\max}}(1 - \hat{\theta}_{n_{\max}})}}{n_{\max}} \right].
 \end{aligned}$$

Note that the random variables $\bar{B}_1, \dots, \bar{B}_{n_{\max}}$ realized by $\bar{b}_1, \dots, \bar{b}_{n_{\max}}$ are not independent because the sample means are computed from the same set of realizations. Also, the random variable, $[\text{CI}]_n$ realized by $[\text{ci}]_n$ are not joint with confidence γ , but in practice this is a computationally efficient way to estimate the parameter. In particular, by plotting $\hat{\theta}_n$ and $[\text{ci}]_n$ for $n = 1, \dots, n_{\max}$, one can gain the convergence behavior of the simulation. In effect, we only perform one experiment, but we interpret it as n_{\max} experiments.

Figure 10.4 shows an example of computing the sample means and confidence intervals in a cumulative manner. The figure confirms that the estimate converges to the true parameter value of $\theta = 0.5$. The confidence interval is a good indicator of the quality of the solution. The error (and the confidence interval) converges at the rate of $n^{-1/2}$, which agrees with the theory.

Chapter 11

Estimation: Areas and Volumes

DRAFT V1.1 © The Authors. License: Creative Commons [BY-NC-SA 3.0](#)

11.1 Calculating an Area

11.1.1 Objective

We are given a two-dimensional domain D in a rectangle $R = [a_1, b_1] \times [a_2, b_2]$. We would like to find, or estimate, the area of D , A_D . Note that the area of the bounding rectangle, $A_R = (b_1 - a_1)(b_2 - a_2)$, is known.

11.1.2 A Continuous Uniform Random Variable

Let $X \equiv (X_1, X_2)$ be a uniform random variable over R . We know that, by the definition of uniform distribution,

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} 1/A_R, & (x_1, x_2) \in R \\ 0, & (x_1, x_2) \notin R \end{cases},$$

and we know how to sample from f_{X_1, X_2} by using independence and univariate uniform densities. Finally, we can express the probability that X takes on a value in D as

$$P(X \in D) = \int_D f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 = \frac{1}{A_R} \int_D dx_1 dx_2 = \frac{A_D}{A_R}.$$

Intuitively, this means that the probability of a random “dart” landing in D is equal to the fraction of the area that D occupies with respect to R .

11.1.3 A Bernoulli Random Variable

Let us introduce a Bernoulli random variable,

$$B = \begin{cases} 1 & X \in D \text{ with probability } \theta \\ 0 & X \notin D \text{ with probability } 1 - \theta \end{cases}.$$

But,

$$P(X \in D) = A_D/A_R,$$

So, by our usual transformation rules,

$$\theta \equiv \frac{A_D}{A_R} .$$

In other words, if we can estimate θ we can estimate $A_D = A_R\theta$. We know how to estimate θ — same as coin flips. But, how do we sample B if we do not know θ ?

11.1.4 Estimation: Monte Carlo

We draw a sample of random vectors,

$$(x_1, x_2)_1, (x_1, x_2)_2, \dots, (x_1, x_2)_n, \dots, (x_1, x_2)_{n_{\max}}$$

and then map the sampled pairs to realization of Bernoulli random variables

$$(x_1, x_2)_i \rightarrow b_i, \quad i = 1, \dots, n_{\max} .$$

Given the realization of Bernoulli random variables,

$$b_1, \dots, b_n, \dots, b_{n_{\max}} ,$$

we can apply the technique discussed in Section 10.3 to compute the sample means and confidence intervals: for $n = 1, \dots, n_{\max}$,

$$\hat{\theta}_n = \bar{b}_n = \frac{1}{n} \sum_{i=1}^n b_i \quad \text{and} \quad [\text{ci}]_n = \left[\hat{\theta}_n - z \frac{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}}{n}, \hat{\theta}_n + z \frac{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}}{n} \right] .$$

Thus, given the mapping from the sampled pairs to Bernoulli variables, we can estimate the parameter.

The only remaining question is how to construct the mapping $(x_1, x_2)_n \rightarrow b_n$, $n = 1, \dots, n_{\max}$. The appropriate mapping is, given $(x_1, x_2)_n \in R$,

$$b_n = \begin{cases} 1, & (x_1, x_2)_n \in D \\ 0, & (x_1, x_2)_n \notin D \end{cases} .$$

To understand why mapping works, we can look at the random variables: given $(X_1, X_2)_n$,

$$B_n = \begin{cases} 1, & (X_1, X_2)_n \in D \\ 0, & (X_1, X_2)_n \notin D \end{cases} .$$

But,

$$P((X_1, X_2)_n \in D) = \theta ,$$

so

$$P(B_n = 1) = \theta ,$$

for $\theta = A_D/A_R$.

This procedure can be intuitively described as

1. Throw n “random darts” at R

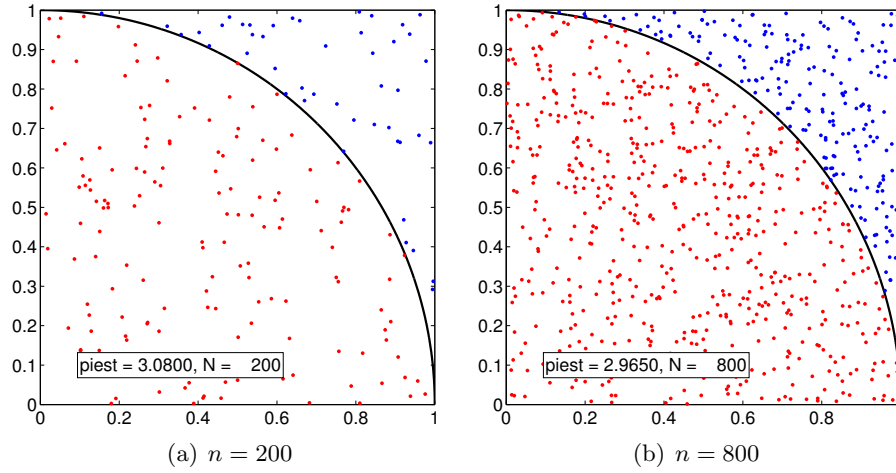


Figure 11.1: Estimation of π by Monte Carlo method.

2. Estimate $\theta = A_D/A_R$ by fraction of darts that land in D

Finally, for A_D , we can develop an estimate

$$(\hat{A}_D)_n = A_R \hat{\theta}_n$$

and confidence interval

$$[ci_{A_D}]_n = A_R [ci]_n .$$

Example 11.1.1 Estimating π by Monte Carlo method

Let us consider an example of estimating the area using Monte Carlo method. In particular, we estimate the area of a circle with unit radius centered at the origin, which has the area of $\pi r^2 = \pi$. Noting the symmetry of the problem, let us estimate the area of the quarter circle in the first quadrant and then multiply the area by four to obtain the estimate of π . In particular, we sample from the square

$$R = [0, 1] \times [0, 1]$$

having the area of $A_R = 1$ and aim to determine the area of the quarter circle D with the area of A_D . Clearly, the analytical answer to this problem is $A_D = \pi/4$. Thus, by estimating A_D using the Monte Carlo method and then multiplying A_D by four, we can estimate the value π .

The sampling procedure is illustrated in Figure 11.1. To determine whether a given sample $(x_1, x_2)_n$ is in the quarter circle, we can compute its distance from the center and determine if the distance is greater than unity, i.e. the Bernoulli variable is assigned according to

$$b_n = \begin{cases} 1, & \overline{x_1^2 + x_2^2} \leq 1 \\ 0, & \text{otherwise} \end{cases} .$$

The samples that evaluates to $b_n = 1$ and 0 are plotted in red and blue, respectively. Because the samples are drawn uniformly from the square, the fraction of red dots is equal to the fraction of the area occupied by the quarter circle. We show in Figure 11.2 the convergence of the Monte Carlo estimation: we observe the anticipated square-root behavior. Note in the remainder of this section we shall use the more conventional N rather than n for sample size.

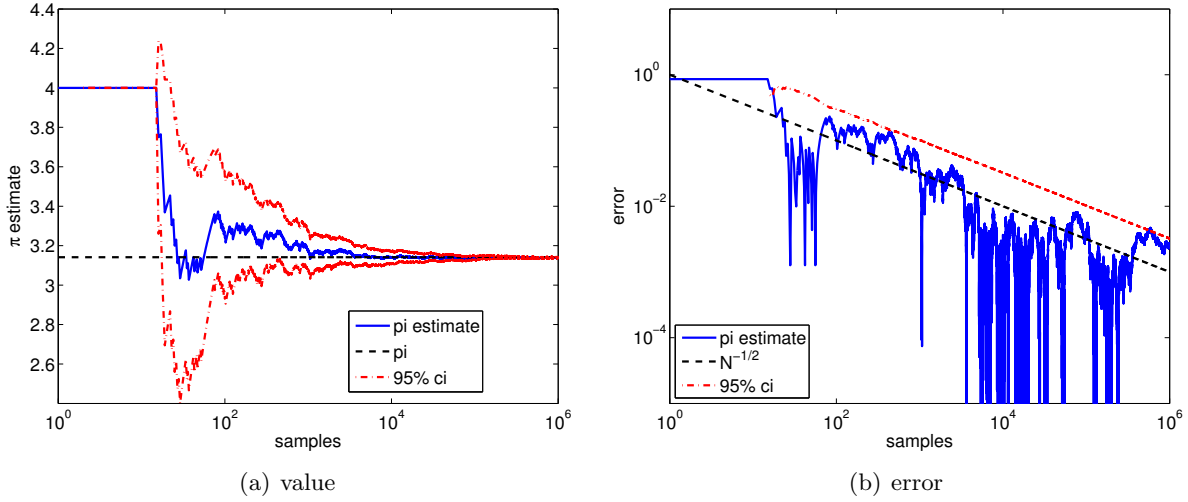


Figure 11.2: Convergence of the π estimate with the number of samples.

11.1.5 Estimation: Riemann Sum

As a comparison, let us also use the midpoint rule to find the area of a two-dimensional region D . We first note that the area of D is equal to the integral of a characteristic function

$$\chi(x_1, x_2) = \begin{cases} 1, & (x_1, x_2) \in D \\ 0, & \text{otherwise} \end{cases},$$

over the domain of integration R that encloses D . For simplicity, we consider rectangular domain $R = [a_1, b_1] \times [a_2, b_2]$. We discretize the domain into $N/2$ little rectangles, each with the width of $(b_1 - a_1)/N/2$ and the height of $(b_2 - a_2)/N/2$. We further divide each rectangle into two right triangles to obtain a triangulation of R . The area of each little triangle K is $A_K = (b_1 - a_1)(b_2 - a_2)/N$. Application of the midpoint rule to integrate the characteristic function yields

$$A_D \approx (\hat{A}_D^{\text{Rie}})_N = \sum_K A_K \chi(x_{c,K}) = \sum_K \frac{(b_1 - a_1)(b_2 - a_2)}{N} \chi(x_{c,K}),$$

where $x_{c,K}$ is the centroid (i.e. the midpoint) of the triangle. Noting that $A_R = (b_1 - a_1)(b_2 - a_2)$ and rearranging the equation, we obtain

$$(\hat{A}_D^{\text{Rie}})_N = A_R \frac{1}{N} \sum_K \chi(x_{c,K}).$$

Because the characteristic function takes on either 0 or 1, the summation is simply equal to the number of times that $x_{c,K}$ is in the region D . Thus, we obtain our final expression

$$\frac{(\hat{A}_D^{\text{Rie}})_N}{A_R} = \frac{\text{number of points in } D}{N}.$$

Note that the expression is very similar to that of the Monte Carlo integration. The main difference is that the sampling points are structured for the Riemann sum (i.e. the centroid of the triangles).

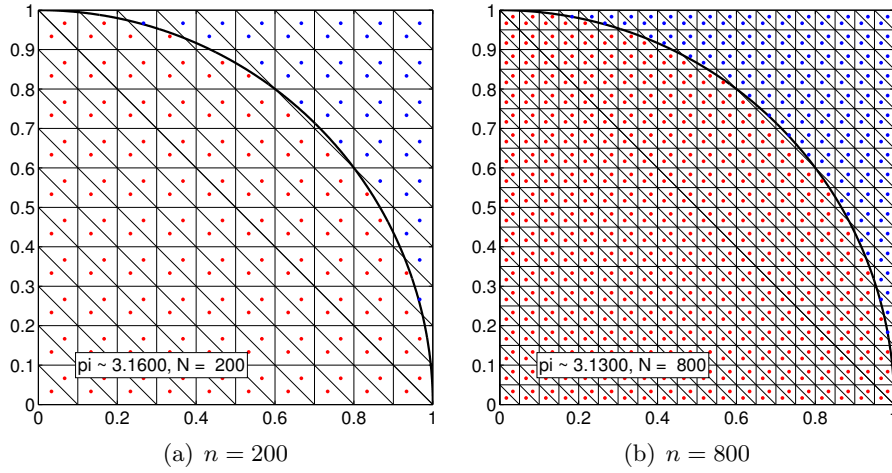


Figure 11.3: Estimation of π by deterministic Riemann sum.

We can also estimate the error incurred in this process. First, we note that we cannot directly apply the error convergence result for the midpoint rule developed previously because the derivation relied on the smoothness of the integrand. The characteristic function is discontinuous along the boundary of D and thus is not smooth. To estimate the error, for simplicity, let us assume the domain size is the same in each dimension, i.e. $a = a_1 = a_2$ and $b = b_1 = b_2$. Then, the area of each square is

$$h^2 = (b - a)^2 / N .$$

Then,

$$(\hat{A}_D^{\text{Rie}})_N = (\text{number of points in } D) \cdot h^2 .$$

There are no errors created by little squares fully inside or fully outside D . All the error is due to the squares that intersect the perimeter. Thus, the error bound can be expressed as

$$\begin{aligned} \text{error} &\approx (\text{number of squares that intersect } D) \cdot h^2 \approx (\text{Perimeter}_D / h) \cdot h^2 \\ &= \mathcal{O}(h) = \mathcal{O}\left(\frac{\overline{A_R}}{N}\right) . \end{aligned}$$

Note that this is an example of *a priori* error estimate. In particular, unlike the error estimate based on the confidence interval of the sample mean for Monte Carlo, this estimate is not constant-free. That is, while we know the asymptotic rate at which the method converges, it does not tell us the actual magnitude of the error, which is problem-dependent. A constant-free estimate typically requires an *a posteriori* error estimate, which incorporates the information gathered about the particular problem of interest. We show in Figure 11.3 the Riemann sum grid, and in Figure 11.4 the convergence of the Riemann sum approach compared to the convergence of the Monte Carlo approach for our π example.

Example 11.1.2 Integration of a rectangular area

In the case of finding the area of a quarter circle, the Riemann sum performed noticeably better than the Monte Carlo method. However, this is not true in general. To demonstrate this, let us consider integrating the area of a rectangle. In particular, we consider the region

$$D = [0.2, 0.7] \times [0.2, 0.8] .$$

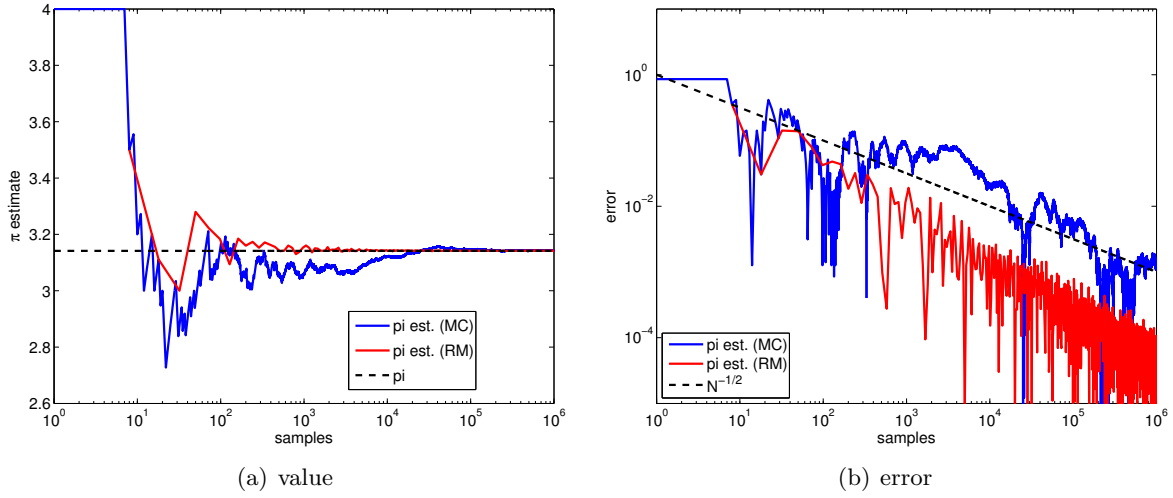


Figure 11.4: Convergence of the π estimate with the number of samples using Riemann sum.

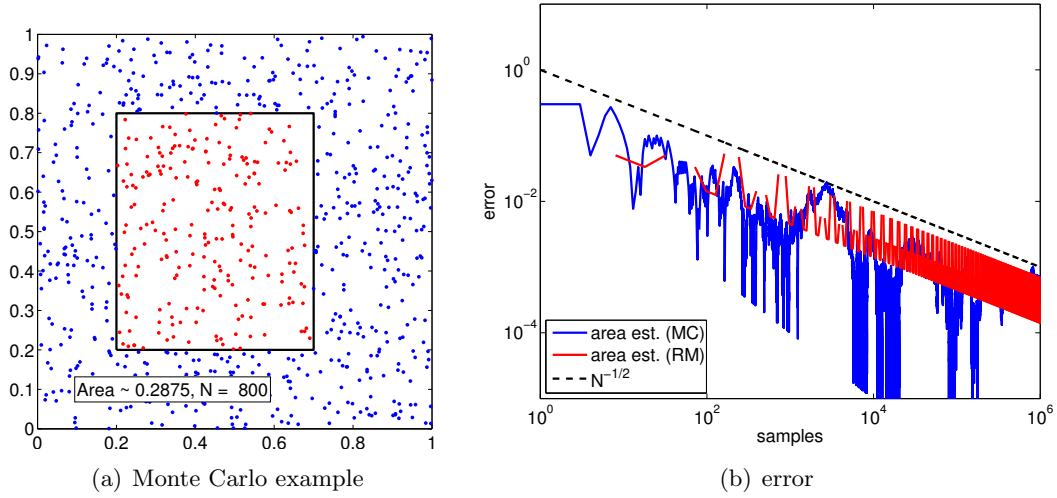


Figure 11.5: The geometry and error convergence for the integration over a rectangular area.

The area of the rectangle is $A_D = (0.7 - 0.2) \cdot (0.8 - 0.2) = 0.3$.

The Monte Carlo integration procedure applied to the rectangular area is illustrated in Figure 11.5(a). The convergence result in Figure 11.5(b) confirms that both Monte Carlo and Riemann sum converge at the rate of $N^{-1/2}$. Moreover, both methods produce the error of similar level for all ranges of the sample size N considered.

Example 11.1.3 Integration of a complex area

Let us consider estimating the area of a more complex region, shown in Figure 11.6(a). The region D is implicitly defined in the polar coordinate as

$$D = (r, \theta) : r \leq \frac{2}{3} + \frac{1}{3} \cos(4\beta\theta), \quad 0 \leq \theta \leq \frac{\pi}{2},$$

where $r = \sqrt{x^2 + y^2}$ and $\tan(\theta) = y/x$. The particular case shown in the figure is for $\beta = 10$. For

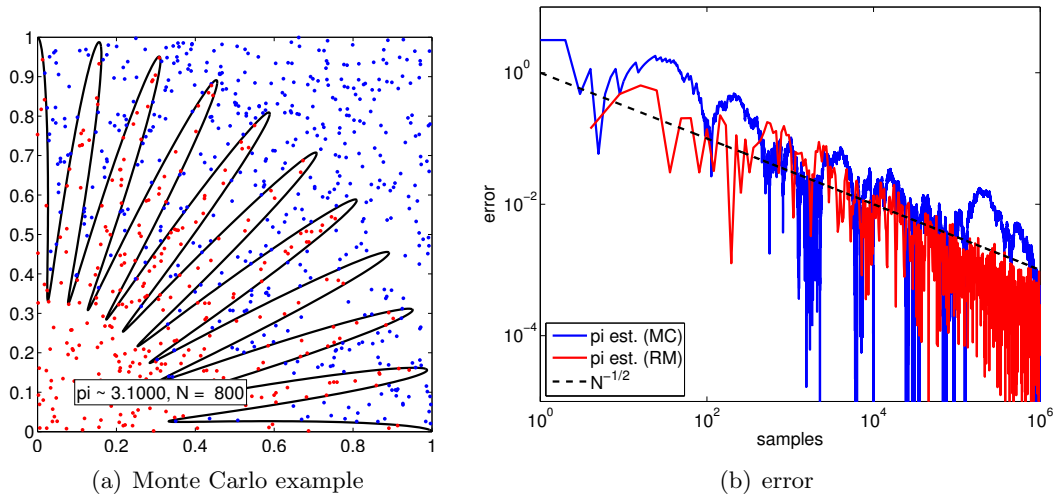


Figure 11.6: The geometry and error convergence for the integration over a more complex area.

any natural number β , the area of the region D is equal to

$$A_D = \int_{\theta=0}^{\pi/2} \int_{r=0}^{2/3+1/3 \cos(4\beta\theta)} r \, dr \, d\theta = \frac{\pi}{8}, \quad \beta = 1, 2, \dots$$

Thus, we can again estimate π by multiplying the estimated area of D by 8.

The result of estimating π by approximating the area of D is shown in Figure 11.6(b). The error convergence plot confirms that both Monte Carlo and Riemann sum converge at the rate of $N^{-1/2}$. In fact, their performances are comparable for this slightly more complex domain.

11.2 Calculation of Volumes in Higher Dimensions

11.2.1 Three Dimensions

Both the Monte Carlo method and Riemann sum used to estimate the area of region D in two dimensions trivially extends to higher dimensions. Let us consider their applications in three dimensions.

Monte Carlo

Now, we sample (X_1, X_2, X_3) uniformly from a parallelepiped $R = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$, where the X_i 's are mutually independent. Then, we assign a Bernoulli random variable according to whether (X_1, X_2, X_3) is inside or outside D as before, i.e.

$$B = \begin{cases} 1, & (X_1, X_2, X_3) \in D \\ 0, & \text{otherwise} \end{cases}.$$

Recall that the convergence of the sample mean to the true value — in particular the convergence of the confidence interval — is related to the Bernoulli random variable and not the X_i 's. Thus, even in three dimensions, we still expect the error to converge as $N^{-1/2}$, where N is the size of the sample.

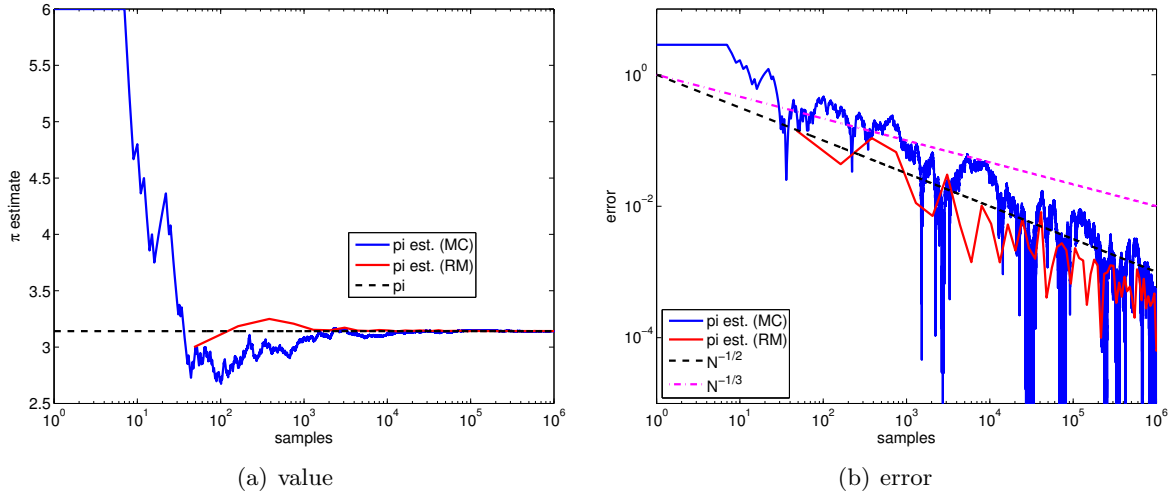


Figure 11.7: Convergence of the π estimate using the volume of a sphere.

Riemann Sum

For simplicity, let us assume the parallelepiped is a cube, i.e. $a = a_1 = a_2 = a_3$ and $b = b_1 = b_2 = b_3$. We consider a grid of N points centered in little cubes of size

$$h^3 = \frac{b - a}{N} ,$$

such that $Nh^3 = V_R$. Extending the two-dimensional case, we estimate the volume of the region D according to

$$\frac{\widehat{V}_D^{\text{Rie}}}{V_R} = \frac{\text{number of points in } D}{N} .$$

However, unlike the Monte Carlo method, the error calculation is dependent on the dimension. The error is given by

$$\begin{aligned} \text{error} &\approx (\text{number of cubes that intersect } D) \cdot h^3 \\ &\approx (\text{surface area of } D/h^2) \cdot h^3 \\ &\approx h \approx N^{-1/3} . \end{aligned}$$

Note that the convergence rate has decreased from $N^{-1/2}$ to $N^{-1/3}$ in going from two to three dimensions.

The result of estimating the area of a sphere in the first octant is shown in Figure 11.7. As expected both the Monte Carlo method and the Riemann sum converge to the correct value. In particular, the Monte Carlo method converges at the expected rate of $N^{-1/2}$. The Riemann sum, on the other hand, converges at a faster rate than the expected rate of $N^{-1/3}$. This superconvergence is due to the symmetry in the tetrahedralization used in integration and the volume of interest. This does not contradict our *a priori* analysis, because the analysis tells us the asymptotic convergence rate for the worst case. The following example shows that the asymptotic convergence rate of the Riemann sum for a general geometry is indeed $N^{-1/2}$.

Example 11.2.1 Integration of a parallelepiped

Let us consider a simpler example of finding the volume of a parallelepiped described by

$$D = [0.1, 0.9] \times [0.2, 0.7] \times [0.1, 0.8] .$$

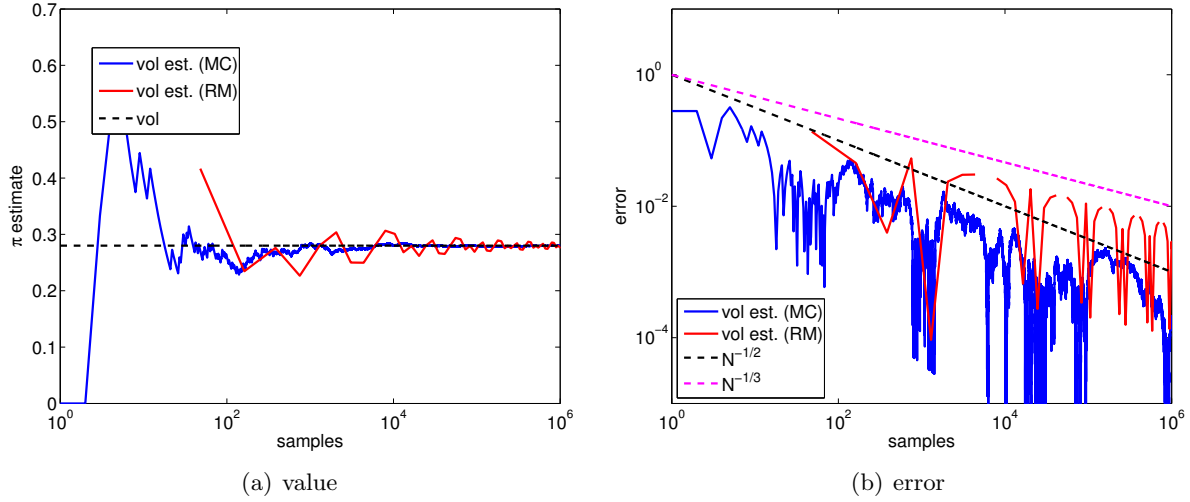


Figure 11.8: Convergence of the π estimate using a complex three-dimensional integration.

The volume of the parallelepiped is $V_D = 0.28$.

Figure 11.8 shows the result of the integration. The figure shows that the convergence rate of the Riemann sum is $N^{-1/3}$, which is consistent with the *a priori* analysis. On the other hand, the Monte Carlo method performs just as well as it did in two dimension, converging at the rate of $N^{-1/2}$. In particular, the Monte Carlo method performs noticeably better than the Riemann sum for large values of N .

Example 11.2.2 Integration of a complex volume

Let us consider a more general geometry, with the domain defined in the spherical coordinate as

$$D = \left\{ (r, \theta, \phi) : r \leq \sin(\theta) \frac{2}{3} + \frac{1}{3} \cos(40\phi) \quad , \quad 0 \leq \theta \leq \frac{\pi}{2}, \quad 0 \leq \phi \leq \frac{\pi}{2} \right\} .$$

The volume of the region is given by

$$V_D = \int_{\phi=0}^{\pi/2} \int_{\theta=0}^{\pi/2} \int_{r=0}^{\sin(\theta)(\frac{2}{3} + \frac{1}{3} \cos(40\phi))} r^2 \sin(\theta) dr d\theta d\phi = \frac{88}{2835} \pi .$$

Thus, we can estimate the value of π by first estimating the volume using Monte Carlo or Riemann sum, and then multiplying the result by $2835/88$.

Figure 11.9 shows the result of performing the integration. The figure shows that the convergence rate of the Riemann sum is $N^{-1/3}$, which is consistent with the *a priori* analysis. On the other hand, the Monte Carlo method performs just as well as it did for the simple sphere case.

11.2.2 General d -Dimensions

Let us generalize our analysis to the case of integrating a general d -dimensional region. In this case, the Monte Carlo method considers a random d -vector, (X_1, \dots, X_d) , and associate with the vector a Bernoulli random variable. The convergence of the Monte Carlo integration is dependent on the Bernoulli random variables and not directly affected by the random vector. In particular, the Monte Carlo method is oblivious to the length of the vector, d , i.e. the dimensionality of the space. Because the standard deviation of the binomial distribution scales as $N^{-1/2}$, we still expect the Monte Carlo method to converge at the rate of $N^{-1/2}$ regardless of d . Thus, Monte Carlo methods

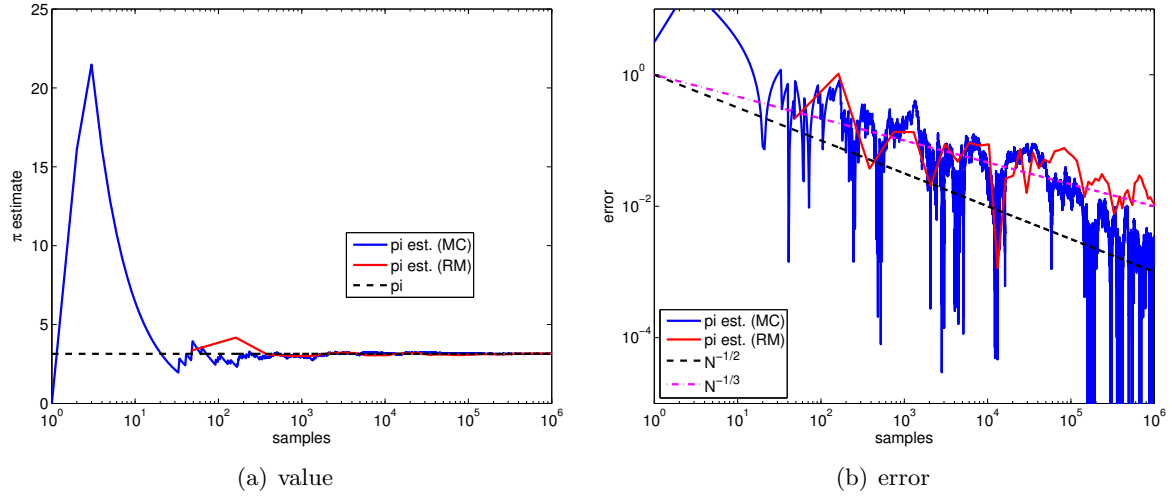


Figure 11.9: Convergence of the π estimate using a complex three-dimensional integration.

do not suffer from so-called curse of dimensionality, in which a method becomes intractable with increase of the dimension of the problem.

On the other hand, the performance of the Riemann sum is a function of the dimension of the space. In a d -dimensional space, each little cube has the volume of N^{-1} , and there are $N^{\frac{d-1}{d}}$ cube that intersect the boundary of D . Thus, the error scales as

$$\text{error} \approx N^{\frac{d-1}{d}} N^{-1} = N^{-1/d} .$$

The convergence rate worsens with the dimension, and this is an example of the curse of dimensionality. While the integration of a physical volume is typically limited to three dimensions, there are many instances in science and engineering where a higher-dimensional integration is required.

Example 11.2.3 integration over a hypersphere

To demonstrate that the convergence of Monte Carlo method is independent of the dimension, let us consider integration of a hypersphere in d -dimensional space. The volume of d -sphere is given by

$$V_D = \frac{\pi^{d/2}}{\Gamma(n/2 + 1)} r^d ,$$

where Γ is the gamma function. We can again use the integration of a d -sphere to estimate the value of π .

The result of estimating the d -dimensional volume is shown in Figure 11.10 for $d = 2, 3, 5, 7$. The error convergence plot shows that the method converges at the rate of $N^{-1/2}$ for all d . The result confirms that Monte Carlo integration is a powerful method for integrating functions in higher-dimensional spaces.

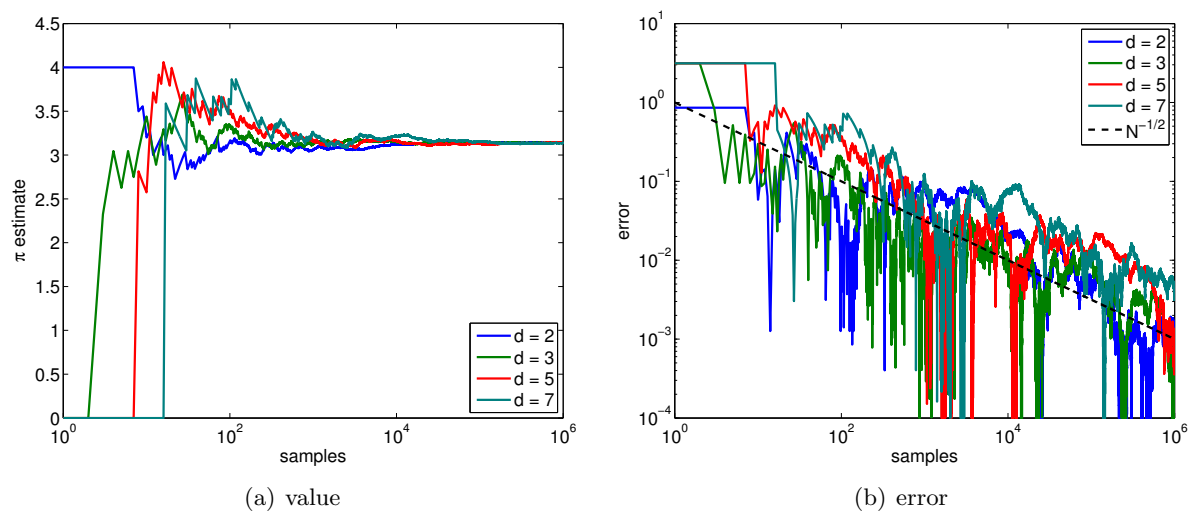


Figure 11.10: Convergence of the π estimate using the Monte Carlo method on d -dimensional hyperspheres.

Chapter 12

Integration: General Procedures

DRAFT V1.1 © The Authors. License: Creative Commons [BY-NC-SA 3.0](#)

This chapter is not available.

This chapter is not available.

Chapter 13

Failure Probabilities

DRAFT V1.1 © The Authors. License: Creative Commons [BY-NC-SA 3.0](#)

13.1 Calculating a Failure Probability

13.1.1 Objective

Let's say there is a set of “environmental” or “load” variables (x_1, x_2, \dots) that affect the performance of an engineering system. For simplicity, let us restrict ourselves to the parameter size of two, so that we only have (x_1, x_2) . We also assume that there are two “performance” metrics, $g_1(x_1, x_2)$ and $g_2(x_1, x_2)$. Without loss of generality, let's assume smaller g_1 and g_2 means better performance (we can always consider negative of the performance variable if larger values imply better performance). In fact, we assume that we wish to confirm that the performance metrics are below certain thresholds, i.e.

$$g_1(x_1, x_2) \leq \tau_1 \quad \text{and} \quad g_2(x_1, x_2) \leq \tau_2 .$$

Equivalently, we wish to avoid failure, which is defined as

$$g_1(x_1, x_2) > \tau_1 \quad \text{or} \quad g_2(x_1, x_2) > \tau_2 .$$

Suppose that (x_1, x_2) reside in some rectangle R . We now choose to interpret (x_1, x_2) as realizations of a random vector $X = (X_1, X_2)$ with prescribed probability density function $f_X(x_1, x_2) = f_{X_1, X_2}(x_1, x_2)$. We then wish to quantify the *failure probability* θ_F , defined by

$$\theta_F = P(g_1(X_1, X_2) > \tau_1 \text{ or } g_2(X_1, X_2) > \tau_2) .$$

We note that g_1 and g_2 are deterministic functions; however, because the argument to the functions are random variables, the output $g_1(X_1, X_2)$ and $g_2(X_1, X_2)$ are random variables. Thus, the failure is described probabilistically. If the bounds on the environmental variables (x_1, x_2) are known *a priori* one could design a system to handle the worst possible cases; however, the system design to handle very rare events may be over designed. Thus, probabilistic approach may be appropriate in many engineering scenarios.

In any probabilistic simulation, we must make sure that the probability density of the random variable, f_X , is meaningful and that the interpretation of the probabilistic statement is relevant. For example, in constructing the distribution, a good estimate may be obtained from statistical data (i.e. by sampling a population). The failure probability θ_F can be interpreted as either (i) probability of failure for the next “random” set of environmental or operating condition, or (ii) frequency of failure over a population.

13.1.2 An Integral

We now show that the computation of failure probability is similar to computing the area. Let us define R to be the region from which $X = (X_1, X_2)$ is sampled (not necessarily uniformly). In other words, R encompasses all possible values that the environmental variable X can take. Let us also define D to be the region whose element $(x_1, x_2) \in D$ would lead to failure, i.e.

$$D \equiv \{(x_1, x_2) : g_1(x_1, x_2) > \tau_1 \quad \text{or} \quad g_2(x_1, x_2) > \tau_2\} .$$

Then, the failure probability can be expressed as an integral

$$\theta_F = \int_D f_X(x_1, x_2) dx_1 dx_2 .$$

This requires a integration over the region D , which can be complicated depending on the failure criterions.

However, we can simplify the integral using the technique previously used to compute the area. Namely, we introduce a failure indicator or characteristic function,

$$\mathbf{1}_F(x_1, x_2) = \begin{cases} 1, & g_1(x_1, x_2) > \tau_1 \quad \text{or} \quad g_2(x_1, x_2) > \tau_2 \\ 0, & \text{otherwise} \end{cases} .$$

Using the failure indicator, we can simplify the integral over D to that over the simpler domain R , i.e.,

$$\theta_F = \int_R \mathbf{1}(x_1, x_2) f_X(x_1, x_2) dx_1 dx_2 .$$

Note that Monte Carlo methods can be used to evaluate any integral in any number of dimensions. The two main approaches are “hit or miss” and “sample mean,” with the latter more efficient. Our case here is a natural example of the sample mean approach, though it also has the flavor of the “hit or miss.” In practice, variance reduction techniques are often applied to improve the convergence.

13.1.3 A Monte Carlo Approach

We can easily develop a Monte Carlo approach if we can reduce our problem to a Bernoulli random variable with parameter θ_F such that

$$B = \begin{cases} 1, & \text{with probability } \theta_F \\ 0, & \text{with probability } 1 - \theta_F \end{cases} .$$

Then, the computation of the failure probability θ_F becomes the estimation of parameter θ_F through sampling (as in the coin flip example). This is easy: we draw a random vector (X_1, X_2) from f_X — for example, uniform or normal — and then define

$$B = \begin{cases} 1, & g_1(X) > \tau_1 \quad \text{or} \quad g_2(X) > \tau_2 \\ 0, & \text{otherwise} \end{cases} . \tag{13.1}$$

Determination of B is easy assuming we can evaluate $g_1(x_1, x_2)$ and $g_2(x_1, x_2)$. But, by definition

$$\begin{aligned} \theta_F &= P(g_1(X) > \tau_1 \quad \text{or} \quad g_2(X) > \tau_2) \\ &= \int_R \mathbf{1}_F(x_1, x_2) f_X(x_1, x_2) dx_1 dx_2 . \end{aligned}$$

So, we draw n_{\max} random variables,

$$(X_1, X_2)_1, (X_1, X_2)_2, \dots, (X_1, X_2)_n, \dots, (X_1, X_2)_{n_{\max}} ,$$

and map them to Bernoulli random variables

$$(X_1, X_2)_n \rightarrow B_n \quad n = 1, \dots, n_{\max} ,$$

according to (13.1). Using this mapping, we assign sample means, $\hat{\Theta}_n$, and confidence intervals, $[\text{CI}_F]_n$, according to

$$\begin{aligned} (\hat{\Theta}_F)_n &= \frac{1}{n} \sum_{i=1}^n B_n \\ [\text{CI}_F]_n &= \left[(\hat{\Theta}_F)_n - z \frac{\overline{(\hat{\Theta}_F)_n(1 - (\hat{\Theta}_F)_n)}}{n}, (\hat{\Theta}_F)_n + z \frac{\overline{(\hat{\Theta}_F)_n(1 - (\hat{\Theta}_F)_n)}}{n} \right] . \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

2.086 Numerical Computation for Mechanical Engineers
Spring 2012

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.