14

Introduction to probability theory

The theory of probability is so useful that it is required in almost every branch of science. In physics, it is of basic importance in quantum mechanics, kinetic theory, and thermal and statistical physics to name just a few topics. In this chapter the reader is introduced to some of the fundamental ideas that make probability theory so useful. We begin with a review of the definitions of probability, a brief discussion of the fundamental laws of probability, and methods of counting (some facts about permutations and combinations), probability distributions are then treated.

A notion that will be used very often in our discussion is 'equally likely'. This cannot be defined in terms of anything simpler, but can be explained and illustrated with simple examples. For example, heads and tails are equally likely results in a spin of a fair coin; the ace of spades and the ace of hearts are equally likely to be drawn from a shuffled deck of 52 cards. Many more examples can be given to illustrate the concept of 'equally likely'.

A definition of probability

Now a question that arises naturally is that of how shall we measure the probability that a particular case (or outcome) in an experiment (such as the throw of dice or the draw of cards) out of many equally likely cases that will occur. Let us flip a coin twice, and ask the question: what is the probability of it coming down heads at least once. There are four equally likely results in flipping a coin twice: *HH*, *HT*, *TH, TT*, where *H* stands for head and *T* for tail. Three of the four results are favorable to at least one head showing, so the probability of getting one head is 3/4. In the example of drawn cards, what is the probability of drawing the ace of spades? Obviously there is one chance out of 52, and the probability, accordingly, is 1/52. On the other hand, the probability of drawing an ace is four times as great $-4/52$, for there are four aces, equally likely. Reasoning in this way, we are led to give the notion of probability the following definition:

> If there are N mutually exclusive, collective exhaustive, and equally likely outcomes of an experiment, and n of these are favorable to an event A, then the probability $p(A)$ of an event A is n/N: $p = n/N$, or

$$p(A) = \frac{\text{number of outcomes favorable to } A}{\text{total number of results}}. \tag{14.1}$$

We have made no attempt to predict the result, just to measure it. The definition of probability given here is often called *a posteriori* probability.

The terms exclusive and exhaustive need some attention. Two events are said to be mutually exclusive if they cannot both occur together in a single trial; and the term collective exhaustive means that all possible outcomes or results are enumerated in the N outcomes.

If an event is certain not to occur its probability is zero, and if an event is certain to occur, then its probability is 1. Now if p is the probability that an event will occur, then the probability that it will fail to occur is $1 - p$, and we denote it by q:

$$q = 1 - p. \tag{14.2}$$

If p is the probability that an event will occur in an experiment, and if the experiment is repeated M times, then the expected number of times the event will occur is Mp. For sufficiently large M, Mp is expected to be close to the actual number of times the event will occur. For example, the probability of a head appearing when tossing a coin is $1/2$, the expected number of times heads appear is $4 \times 1/2$ or 2. Actually, heads will not always appear twice when a coin is tossed four times. But if it is tossed 50 times, the number of heads that appear will, on the average, be close to 25 ($50 \times 1/2 = 25$). Note that closeness is computed on a percentage basis: 20 is 20% of 25 away from 25 while 1 is 50% of 2 away from 2.

Sample space

The equally likely cases associated with an experiment represent the possible outcomes. For example, the 36 equally likely cases associated with the throw of a pair of dice are the 36 ways the dice may fall, and if 3 coins are tossed, there are 8 equally likely cases corresponding to the 8 possible outcomes. A list or set that consists of all possible outcomes of an experiment is called a sample space and each individual outcome is called a sample point (a point of the sample space). The outcomes composing the sample space are required to be mutually exclusive. As an example, when tossing a die the outcomes 'an even number shows' and 'number 4 shows' cannot be in the same sample space. Often there will be more than one sample space that can describe the outcome of an experiment but there is usually only one that will provide the most information. In a throw of a fair die, one sample space is the set of all possible outcomes {1, 2, 3, 4, 5, 6}, and another could be {even} or {odd}.

A finite sample space is one that has only a finite number of points. The points of the sample space are weighted according to their probabilities. To see this, let the points have the probabilities

$$p_1, p_2, \ldots, p_N$$

with

$$p_1 + p_2 + \cdots + p_N = 1.$$

Suppose the first n sample points are favorable to another event A. Then the probability of A is defined to be

$$p(A) = p_1 + p_2 + \cdots + p_n.$$

Thus the points of the sample space are weighted according to their probabilities. If each point has the sample probability $1/n$, then $p(A)$ becomes

$$p(A) = \frac{1}{N} + \frac{1}{N} + \cdots + \frac{1}{N} = \frac{n}{N}$$

and this definition is consistent with that given by Eq. (14.1).

A sample space with constant probability is called uniform. Non-uniform sample spaces are more common. As an example, let us toss four coins and count the number of heads. An appropriate sample space is composed of the outcomes

0 heads, 1 head, 2 heads, 3 heads, 4 heads,

with respective probabilities, or weights

$$1/16, \quad 4/16, \quad 6/16, \quad 4/16, \quad 1/16.$$

The four coins can fall in $2 \times 2 \times 2 \times 2 = 2^4$, or 16 ways. They give no heads (all land tails) in only one outcome, and hence the required probability is 1/16. There are four ways to obtain 1 head: a head on the first coin or on the second coin, and so on. This gives 4/16. Similarly we can obtain the probabilities for the other cases.

We can also use this simple example to illustrate the use of sample space. What is the probability of getting at least two heads? Note that the last three sample points are favorable to this event, hence the required probability is given by

$$\frac{6}{16} + \frac{4}{16} + \frac{1}{16} = \frac{11}{16}.$$

Methods of counting

In many applications the total number of elements in a sample space or in an event needs to be counted. A fundamental principle of counting is this: if one thing can be done in n different ways and another thing can be done in m different ways, then both things can be done together or in succession in mn different ways. As an example, in the example of throwing a pair of dice cited above, there are 36 equally like outcomes: the first die can fall in six ways, and for each of these the second die can also fall in six ways. The total number of ways is

$$6 + 6 + 6 + 6 + 6 + 6 = 6 \times 6 = 36$$

and these are equally likely.

Enumeration of outcomes can become a lengthy process, or it can become a practical impossibility. For example, the throw of four dice generates a sample space with $6^4 = 1296$ elements. Some systematic methods for the counting are desirable. Permutation and combination formulas are often very useful.

Permutations

A permutation is a particular ordered selection. Suppose there are n objects and r of these objects are arranged into r numbered spaces. Since there are n ways of choosing the first object, and after this is done there are $n - 1$ ways of choosing the second object, ..., and finally $n - (r - 1)$ ways of choosing the rth object, it follows by the fundamental principle of counting that the number of different arrangements or permutations is given by

$$_nP_r = n(n-1)(n-2)\cdots(n-r+1). \tag{14.3}$$

where the product on the right-hand side has r factors. We call $_nP_r$ the number of permutations of n objects taken r at a time. When $r = n$, we have

$$_nP_n = n(n-1)(n-2)\cdots 1 = n!.$$

We can rewrite $_nP_r$ in terms of factorials:

$$\begin{aligned} _nP_r &= n(n-1)(n-2)\cdots(n-r+1) \\ &= n(n-1)(n-2)\cdots(n-r+1)\frac{(n-r)\cdots 2 \times 1}{(n-r)\cdots 2 \times 1} \\ &= \frac{n!}{(n-r)!}. \end{aligned}$$

When $r = n$, we have $_nP_n = n!/(n-n)! = n!/0!$. This reduces to $n!$ if we have $0! = 1$ and mathematicians actually take this as the definition of $0!$.

Suppose the n objects are not all different. Instead, there are n_1 objects of one kind (that is, indistinguishable from each other), n_2 that is of a second kind, ..., n_k

of a kth kind so that $n_1 + n_2 + \cdots + n_k = n$. A natural question is that of how many distinguishable arrangements are there of these n objects. Assuming that there are N different arrangements, and each distinguishable arrangement appears $n_1!$, $n_2!, \ldots$ times, where $n_1!$ is the number of ways of arranging the n_1 objects, similarly for $n_2!, \ldots, n_k!$. Then multiplying N by $n_1!n_2!, \ldots, n_k!$ we obtain the number of ways of arranging the n objects if they were all distinguishable, that is, ${}_nP_n = n!$:

$$Nn_1!n_2! \cdots n_k! = n! \qquad \text{or} \qquad N = n!/(n_1!n_2! \ldots n_k!).$$

N is often written as ${}_nP_{n_1 n_2 \ldots n_k}$, and then we have

$${}_nP_{n_1 n_2 \ldots n_k} = \frac{n!}{n_1!n_2! \cdots n_k!}. \tag{14.4}$$

For example, given six coins: one penny, two nickels and three dimes, the number of permutations of these six coins is

$${}_6P_{123} = 6!/1!2!3! = 60.$$

Combinations

A permutation is a particular ordered selection. Thus 123 is a different permutation from 231. In many problems we are interested only in selecting objects without regard to order. Such selections are called combinations. Thus 123 and 231 are now the same combination. The notation for a combination is ${}_nC_r$ which means the number of ways in which r objects can be selected from n objects without regard to order (also called the combination of n objects taken r at a time). Among the ${}_nP_r$ permutations there are $r!$ that give the same combination. Thus, the total number of permutations of n different objects selected r at a time is

$$r!\,{}_nC_r = {}_nP_r = \frac{n!}{(n-r)!}.$$

Hence, it follows that

$${}_nC_r = \frac{n!}{r!(n-r)!}. \tag{14.5}$$

It is straightforward to show that

$${}_nC_r = \frac{n!}{r!(n-r)!} = \frac{n!}{[n-(n-r)]!(n-r)!} = {}_nC_{n-r}.$$

${}_nC_r$ is often written as

$${}_nC_r = \binom{n}{r}.$$

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

The numbers (14.5) are often called binomial coefficients because they arise in the binomial expansion

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n}y^n.$$

When n is very large a direct evaluation of $n!$ is impractical. In such cases we use Stirling's approximate formula

$$n! \approx \sqrt{2\pi n}n^n e^{-n}.$$

The ratio of the left hand side to the right hand side approaches 1 as $n \to \infty$. For this reason the right hand side is often called an asymptotic expansion of the left hand side.

Fundamental probability theorems

So far we have calculated probabilities by directly making use of the definitions; it is doable but it is not always easy. Some important properties of probabilities will help us to cut short our computation works. These important properties are often described in the form of theorems. To present these important theorems, let us consider an experiment, involving two events A and B, with N equally likely outcomes and let

$$n_1 = \text{number of outcomes in which } A \text{ occurs, but not } B,$$

$$n_2 = \text{number of outcomes in which } B \text{ occurs, but not } A,$$

$$n_3 = \text{number of outcomes in which both } A \text{ and } B \text{ occur},$$

$$n_4 = \text{number of outcomes in which neither } A \text{ nor } B \text{ occurs}.$$

This covers all possibilities, hence $n_1 + n_2 + n_3 + n_4 = N$.

The probabilities of A and B occurring are respectively given by

$$P(A) = \frac{n_1 + n_3}{N}, \qquad P(B) = \frac{n_2 + n_3}{N}, \tag{14.6}$$

the probability of either A or B (or both) occurring is

$$P(A+B) = \frac{n_1 + n_2 + n_3}{N}, \tag{14.7}$$

and the probability of both A and B occurring successively is

$$P(AB) = \frac{n_3}{N}. \tag{14.8}$$

Let us rewrite $P(AB)$ as

$$P(AB) = \frac{n_3}{N} = \frac{n_1 + n_3}{N}\frac{n_3}{n_1 + n_3}.$$

Now $(n_1 + n_3)/N$ is $P(A)$ by definition. After A has occurred, the only possible cases are the $(n_1 + n_3)$ cases favorable to A. Of these, there are n_3 cases favorable to B, the quotient $n_3/(n_1 + n_3)$ represents the probability of B when it is known that A occurred, $PA(B)$. Thus we have

$$P(AB) = P(A)P_A(B). \tag{14.9}$$

This is often known as the theorem of joint (or compound) probability. In words, the joint probability (or the compound probability) of A and B is the product of the probability that A will occur times the probability that B will occur if A does. $P_A(B)$ is called the conditional probability of B given A (that is, given that A has occurred).

To illustrate the theorem of joint probability (14.9), we consider the probability of drawing two kings in succession from a shuffled deck of 52 playing cards. The probability of drawing a king on the first draw is 4/52. After the first king has been drawn, the probability of drawing another king from the remaining 51 cards is 3/51, so that the probability of two kings is

$$\frac{4}{52} \times \frac{3}{51} = \frac{1}{221}.$$

If the events A and B are independent, that is, the information that A has occurred does not influence the probability of B, then $P_A(B) = P(B)$ and the joint probability takes the form

$$P(AB) = P(A)P(B), \text{ for independent events.} \tag{14.10}$$

As a simple example, let us toss a coin and a die, and let A be the event 'head shows' and B is the event '4 shows.' These events are independent, and hence the probability that 4 and a head both show is

$$P(AB) = P(A)P(B) = (1/2)(1/6) = 1/12.$$

Theorem (14.10) can be easily extended to any number of independent events $A, B, C, \ldots$.

Besides the theorem of joint probability, there is a second fundamental relationship, known as the theorem of total probability. To present this theorem, let us go back to Eq. (14.4) and rewrite it in a slightly different form

$$\begin{aligned} P(A+B) &= \frac{n_1 + n_2 + n_3}{N} \\ &= \frac{n_1 + n_2 + 2n_3 - n_3}{N} = \frac{(n_1 + n_3) + (n_2 + n_3) - n_3}{N} \\ &= \frac{n_1 + n_3}{N} + \frac{n_2 + n_3}{N} - \frac{n_3}{N} = P(A) + P(B) - P(AB), \end{aligned}$$

$$P(A+B) = P(A) + P(B) - P(AB). \tag{14.11}$$

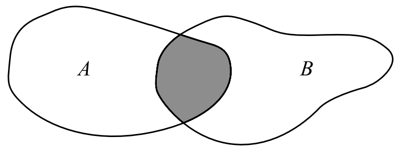

Figure 14.1.

This theorem can be represented diagrammatically by the intersecting points sets A and B shown in Fig. 14.1. To illustrate this theorem, consider the simple example of tossing two dice and find the probability that at least one die gives 2. The probability that both give 2 is 1/36. The probability that the first die gives 2 is 1/6, and similarly for the second die. So the probability that at least one gives 2 is

$$P(A+B) = 1/6 + 1/6 - 1/36 = 11/36.$$

For mutually exclusive events, that is, for events A, B which cannot both occur, $P(AB) = 0$ and the theorem of total probability becomes

$$P(A+B) = P(A) + P(B), \qquad \text{for mutually exclusive events.} \tag{4.12}$$

For example, in the toss of a die, '4 shows' (event A) and '5 shows' (event B) are mutually exclusive, the probability of getting either 4 or 5 is

$$P(A+B) = P(A) + P(B) = 1/6 + 1/6 = 1/3.$$

The theorems of total and joint probability for uniform sample spaces established above are also valid for arbitrary sample spaces. Let us consider a finite sample space, its events E_i are so numbered that $E_1, E_2, \ldots, E_j$ are favorable to $A, E_{j+1}, \ldots, E_k$ are favorable to both A and B, and $E_{k+1}, \ldots, E_m$ are favorable to B only. If the associated probabilities are p_i, then Eq. (14.11) is equivalent to the identity

$$\begin{aligned} p_1 + \cdots + p_m = {} & (p_1 + \cdots + p_j + p_{j+1} + \cdots + p_k) \\ & + (p_{j+1} + \cdots + p_k + p_{k+1} + \cdots + p_m) - (p_{j+1} + \cdots + p_m). \end{aligned}$$

The sums within the three parentheses on the right hand side represent, respectively, $P(A)P(B)$, and $P(AB)$ by definition. Similarly, we have

$$\begin{aligned} P(AB) &= p_{j+1} + \cdots + p_k \\ &= (p_1 + \cdots + p_k)\left(\frac{p_{j+1}}{p_1 + \cdots + p_k} + \cdots + \frac{p_k}{p_1 + \cdots + p_k}\right) \\ &= P(A)P_A(B), \end{aligned}$$

which is Eq. (14.9).

Random variables and probability distributions

As demonstrated above, simple probabilities can be computed from elementary considerations. We need more efficient ways to deal with probabilities of whole classes of events. For this purpose we now introduce the concepts of random variables and a probability distribution.

Random variables

A process such as spinning a coin or tossing a die is called random since it is impossible to predict the final outcome from the initial state. The outcomes of a random process are certain numerically valued variables that are often called random variables. For example, suppose that three dimes are tossed at the same time and we ask how many heads appear. The answer will be 0, 1, 2, or 3 heads, and the sample space S has 8 elements:

$$S = \{TTT, HTT, THT, TTH, HHT, HTH, THH, HHH\}.$$

The random variable X in this case is the number of heads obtained and it assumes the values

$$0, 1, 1, 1, 2, 2, 2, 3.$$

For instance, $X = 1$ corresponds to each of the three outcomes: HTT, THT, TTH. That is, the random variable X can be thought of as a function of the number of heads appear.

A random variable that takes on a finite or countable infinite number of values (that is it has as many values as the natural numbers $1, 2, 3, \ldots$) is called a discrete random variable while one that takes on a non-countable infinite number of values is called a non-discrete or continuous random variable.

Probability distributions

A random variable, as illustrated by the simple example of tossing three dimes at the same time, is a numerical-valued function defined on a sample space. In symbols,

$$X(s_i) = x_i \qquad i = 1, 2, \ldots, n, \tag{14.13}$$

where s_i are the elements of the sample space and x_i are the values of the random variable X. The set of numbers x_i can be finite or infinite.

In terms of a random variable we will write $P(X = x_i)$ as the probability that the random variable X takes the value x_i, and $P(X < x_i)$ as the probability that the random variable takes values less than x_i, and so on. For simplicity, we often write $P(X = x_i)$ as p_i. The pairs (x_i, p_i) for $i = 1, 2, 3, \ldots$ define the probability

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

distribution or probability function for the random variable X. Evidently any probability distribution p_i for a discrete random variable must satisfy the following conditions:

(*i*) $0 \le p_i \le 1$;
(*ii*) the sum of all the probabilities must be unity (certainty), $\sum_i p_i = 1$.

Expectation and variance

The expectation or expected value or mean of a random variable is defined in terms of a weighted average of outcomes, where the weighting is equal to the probability p_i with which x_i occurs. That is, if X is a random variable that can take the values $x_1, x_2, \ldots,$ with probabilities $p_1, p_2, \ldots,$ then the expectation or expected value $E(X)$ is defined by

$$E(X) = p_1x_1 + p_2x_2 + \cdots = \sum_i p_i x_i. \tag{14.14}$$

Some authors prefer to use the symbol μ for the expectation value $E(X)$. For the three dimes tossed at the same time, we have

$$\begin{array}{lccc} x_i = 0 & 1 & 2 & 3 \\ p_i = 1/8 & 3/8 & 3/8 & 1/8 \end{array}$$

and

$$E(X) = \frac{1}{8} \times 0 + \frac{3}{8} \times 1 + \frac{3}{8} \times 2 + \frac{1}{8} \times 3 = \frac{3}{2}.$$

We often want to know how much the individual outcomes are scattered away from the mean. A quantity measure of the spread is the difference $X - E(X)$ and this is called the deviation or residual. But the expectation value of the deviations is always zero:

$$\begin{aligned} E(X - E(X)) &= \sum_i (x_i - E(X))p_i = \sum_i x_i p_i - E(X) \sum_i p_i \\ &= E(X) - E(X) \cdot 1 = 0. \end{aligned}$$

This should not be particularly surprising; some of the deviations are positive, and some are negative, and so the mean of the deviations is zero. This means that the mean of the deviations is not very useful as a measure of spread. We get around the problem of handling the negative deviations by squaring each deviation, thereby obtaining a quantity that is always positive. Its expectation value is called the variance of the set of observations and is denoted by σ^2

$$\sigma^2 = E[(X - E(X))^2] = E[(X - \mu)^2]. \tag{14.15}$$

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

The square root of the variance, σ, is known as the standard deviation, and it is always positive.

We now state some basic rules for expected values. The proofs can be found in any standard textbook on probability and statistics. In the following c is a constant, X and Y are random variables, and $h(X)$ is a function of X:

(1) $E(cX) = cE(X)$;
(2) $E(X + Y) = E(X) + E(Y)$;
(3) $E(XY) = E(X)E(Y)$ (provided X and Y are independent);
(4) $E(h(X)) = \sum_i h(x_i)p_i$ (for a finite distribution).

Special probability distributions

We now consider some special probability distributions in which we will use all the things we have learned so far about probability.

The binomial distribution

Before we discuss the binomial distribution, let us introduce a term, the Bernoulli trials. Consider an experiment such as spinning a coin or throw a die repeatedly. Each spin or toss is called a trial. In any single trial there will be a probability p associated with a particular event (or outcome). If p is constant throughout (that is, does not change from one trial to the next), such trials are then said to be independent and are known as Bernoulli trials.

Now suppose that we have n independent events of some kind (such as tossing a coin or die), each of which has a probability p of success and probability of $q = (1 - p)$ of failure. What is the probability that exactly m of the events will succeed? If we select m events from n, the probability that these m will succeed and all the rest $(n - m)$ will fail is $p^m q^{n-m}$. We have considered only one particular group or combination of m events. How many combinations of m events can be chosen from n? It is the number of combinations of n things taken m at a time: ${}_nC_m$. Thus the probability that exactly m events will succeed from a group of n is

$$f(m) = P(X = m) = {}_nC_m p^m q^{(n-m)}$$
$$= \frac{n!}{m!(n-m)!} p^m q^{(n-m)}. \qquad (14.16)$$

This discrete probability function (14.16) is called the binomial distribution for X, the random variable of the number of successes in the n trials. It gives the probability of exactly m successes in n independent trials with constant probability p. Since many statistical studies involve repeated trials, the binomial distribution has great practical importance.

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

Why is the discrete probability function (14.16) called the binomial distribution? Since for $m = 0, 1, 2, \ldots, n$ it corresponds to successive terms in the binomial expansion

$$(q+p)^n = q^n + {}_nC_1 q^{n-1}p + {}_nC_2 q^{n-2}p^2 + \cdots + p^n = \sum_{m=0}^{n} {}_nC_m p^m q^{n-m}.$$

To illustrate the use of the binomial distribution (14.16), let us find the probability that a one will appear exactly 4 times if a die is thrown 10 times. Here $n = 10$, $m = 4$, $p = 1/6$, and $q = (1 - p) = 5/6$. Hence the probability is

$$f(4) = P(X = 4) = \frac{10!}{4!6!}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6 = 0.0543.$$

A few examples of binomial distributions, computed from Eq. (14.16), are shown in Figs. 14.2, and 14.3 by means of histograms.

One of the key requirements for a probability distribution is that

$$\sum_{m=o}^{n} f(m) = \sum_{m=o}^{n} {}_nC_m p^m q^{n-m} = 1. \tag{14.17}$$

To show that this is in fact the case, we note that

$$\sum_{m=o}^{n} {}_nC_m p^m q^{n-m}$$

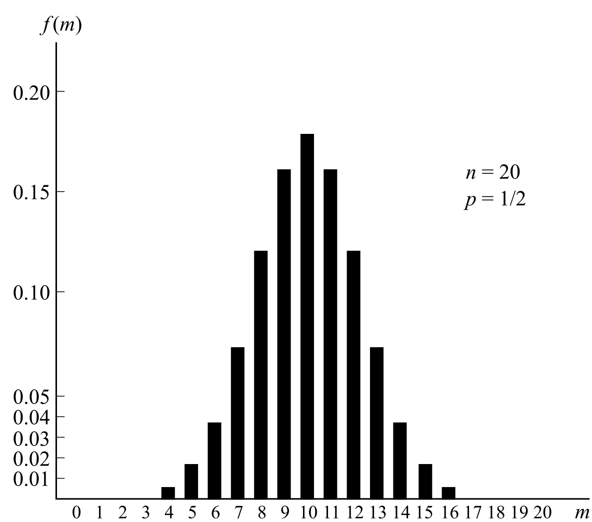

Figure 14.2. The distribution is symmetric about $m = 10$.

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

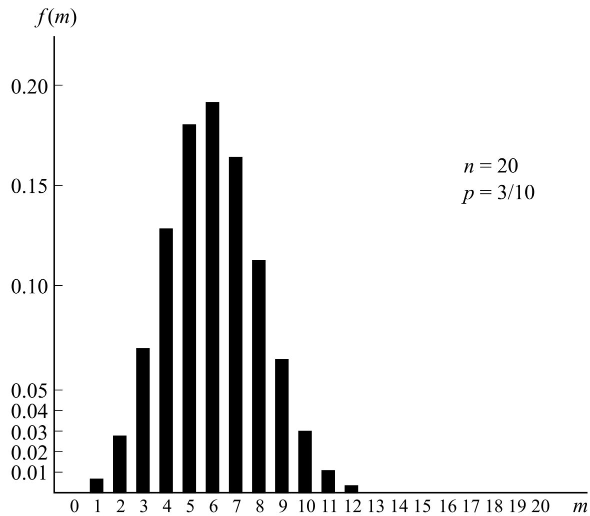

Figure 14.3. The distribution favors smaller value of m.

is exactly equal to the binomial expansion of $(q+p)^n$. But here $q+p=1$, so $(q+p)^n=1$ and our proof is established.

The mean (or average) number of successes, $\bar{m}$, is given by

$$\bar{m}=\sum_{m=0}^{n} m_n C_m p^m (1-p)^{n-m}. \tag{14.18}$$

The sum ranges from $m=0$ to n because in every one of the sets of trials the same number of successes between 0 and n must occur. It is similar to Eq. (14.17); the difference is that the sum in Eq. (14.18) contains an extra factor n. But we can convert it into the form of the sum in Eq. (14.17). Differentiating both sides of Eq. (14.17) with respect to p, which is legitimate as the equation is true for all p between 0 and 1, gives

$$\sum {}_nC_m[mp^{m-1}(1-p)^{n-m}-(n-m)p^m(1-p)^{n-m-1}]=0,$$

where we have dropped the limits on the sum, remembering that m ranges from 0 to n. The last equation can be rewritten as

$$\begin{aligned}\sum m_n C_m p^{m-1}(1-p)^{n-m} &= \sum (n-m)_n C_m p^m (1-p)^{n-m-1}\\ &= n\sum {}_nC_m p^m(1-p)^{n-m-1}-\sum m_n C_m p^m(1-p)^{n-m-1}\end{aligned}$$

or

$$\sum m_n C_m[p^{m-1}(1-p)^{n-m}+p^m(1-p)^{n-m-1}]=n\sum {}_nC_m p^m(1-p)^{n-m-1}.$$

Now multiplying both sides by $p(1-p)$ we get

$$\sum m_n C_m[(1-p)p^m(1-p)^{n-m} + p^{m+1}(1-p)^{n-m}] = np\sum {}_nC_m p^m(1-p)^{n-m}.$$

Combining the two terms on the left hand side, and using Eq. (14.17) in the right hand side we have

$$\sum m_n C_m p^m(1-p)^{n-m} = \sum mf(m) = np. \qquad (14.19)$$

Note that the left hand side is just our original expression for $\bar{m}$, Eq. (14.18). Thus we conclude that

$$\bar{m} = np \qquad (14.20)$$

for the binomial distribution.

The variance σ^2 is given by

$$\sigma^2 = \sum (m-\bar{m})^2 f(m) = \sum (m-np)^2 f(m); \qquad (14.21)$$

here we again drop the summation limits for convenience. To evaluate this sum we first rewrite Eq. (14.21) as

$$\begin{aligned}\sigma^2 &= \sum (m^2 - 2mnp + n^2p^2) f(m)\\ &= \sum m^2 f(m) - 2np\sum mf(m) + n^2p^2\sum f(m).\end{aligned}$$

This reduces to, with the help of Eqs. (14.17) and (14.19),

$$\sigma^2 = \sum m^2 f(m) - (np)^2. \qquad (14.22)$$

To evaluate the first term on the right hand side, we first differentiate Eq. (14.19):

$$\sum m_n C_m[mp^{m-1}(1-p)^{n-m} - (n-m)p^m(1-p)^{n-m-1}] = p,$$

then multiplying by $p(1-p)$ and rearranging terms as before

$$\sum m^2{}_nC_m p^m(1-p)^{n-m} - np\sum m_n C_m p^m(1-p)^{n-m} = np(1-p).$$

By using Eq. (14.19) we can simplify the second term on the left hand side and obtain

$$\sum m^2{}_nC_m p^m(1-p)^{n-m} = (np)^2 + np(1-p)$$

or

$$\sum m^2 f(m) = np(1-p+np).$$

Inserting this result back into Eq. (14.22), we obtain

$$\sigma^2 = np(1-p+np) - (np)^2 = np(1-p) = npq, \qquad (14.23)$$

and the standard deviation σ;

$$\sigma = \sqrt{npq}. \tag{14.24}$$

Two different limits of the binomial distribution for large n are of practical importance: (1) $n \to \infty$ and $p \to 0$ in such a way that the product $np = \lambda$ remains constant; (2) both n and pn are large. The first case will result a new distribution, the Poisson distribution, and the second cases gives us the Gaussian (or Laplace) distribution.

The Poisson distribution

Now $np = \lambda$, so $p = \lambda/n$. The binomial distribution (14.16) then becomes

$$\begin{aligned} f(m) = P(X = m) &= \frac{n!}{m!(n-m)!}\left(\frac{\lambda}{n}\right)^m \left(1 - \frac{\lambda}{n}\right)^{n-m} \\ &= \frac{n(n-1)(n-2)\cdots(n-m+1)}{m!n^m}\lambda^m \left(1 - \frac{\lambda}{n}\right)^{n-m} \\ &= \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{m-1}{n}\right)\frac{\lambda^m}{m!}\left(1 - \frac{\lambda}{n}\right)^{n-m}. \end{aligned} \tag{14.25}$$

Now as $n \to \infty$,

$$\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{m-1}{n}\right) \to 1,$$

while

$$\left(1 - \frac{\lambda}{n}\right)^{n-m} = \left(1 - \frac{\lambda}{n}\right)^{n}\left(1 - \frac{\lambda}{n}\right)^{-m} \to \left(e^{-\lambda}\right)(1) = e^{-\lambda},$$

where we have made use of the result

$$\lim_{n\to\infty}\left(1 + \frac{\alpha}{n}\right)^n = e^{\alpha}.$$

It follows that Eq. (14.25) becomes

$$f(m) = P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!}. \tag{14.26}$$

This is known as the Poisson distribution. Note that $\sum_{m=0}^{\infty} P(X = m) = 1$, as it should.

The Poisson distribution has the mean

$$E(X) = \sum_{m=0}^{\infty} \frac{m\lambda^m e^{-\lambda}}{m!} = \sum_{m=1}^{\infty} \frac{\lambda^m e^{-\lambda}}{(m-1)!} = \lambda \sum_{m=0}^{\infty} \frac{\lambda^m e^{-\lambda}}{m!}$$

$$= \lambda e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = \lambda e^{-\lambda} e^{\lambda} = \lambda, \tag{14.27}$$

where we have made use of the result

$$\sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{\lambda}.$$

The variance σ^2 of the Poisson distribution is

$$\sigma^2 = \mathrm{Var}(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$$

$$= \sum_{m=0}^{\infty} \frac{m^2 \lambda^m e^{-\lambda}}{m!} - \lambda^2 = e^{-\lambda} \sum_{m=1}^{\infty} \frac{m\lambda^m}{(m-1)!} - \lambda^2$$

$$= e^{-\lambda} \lambda \frac{d}{d\lambda} \left(\lambda e^{\lambda}\right) - \lambda^2 = \lambda. \tag{14.28}$$

To illustrate the use of the Poisson distribution, let us consider a simple example. Suppose the probability that an individual suffers a bad reaction from a flu injection is 0.001; what is the probability that out of 2000 individuals (*a*) exactly 3, (*b*) more than 2 individuals will suffer a bad reaction? Now X denotes the number of individuals who suffer a bad reaction and it is binomially distributed. However, we can use the Poisson approximation, because the bad reactions are assumed to be rare events. Thus

$$P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!}, \quad \text{with } \lambda = mp = (2000)(0.001) = 2:$$

(*a*) $P(X = 3) = \dfrac{2^3 e^{-2}}{3!} = 0.18;$

(*b*) $$P(X > 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - \left[\frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} + \frac{2^2 e^{-2}}{2!}\right]$$

$$= 1 - 5e^{-2} = 0.323.$$

An exact evaluation of the probabilities using the binomial distribution would require much more labor.

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

The Poisson distribution is very important in nuclear physics. Suppose that we have n radioactive nuclei and the probability for any one of these to decay in a given interval of time T is p, then the probability that m nuclei will decay in the interval T is given by the binomial distribution. However, n may be a very large number (such as 10^{23}), and p may be the order of 10^{-20}, and it is impractical to evaluate the binomial distribution with numbers of these magnitudes. Fortunately, the Poisson distribution can come to our rescue.

The Poisson distribution has its own significance beyond its connection with the binomial distribution and it can be derived mathematically from elementary considerations. In general, the Poisson distribution applies when a very large number of experiments is carried out, but the probability of success in each is very small, so that the expected number of successes is a finite number.

The Gaussian (or normal) distribution

The second limit of the binomial distribution that is of interest to us results when both n and pn are large. Clearly, we assume that m, n, and $n-m$ are large enough to permit the use of Stirling's formula ($n! \approx \sqrt{2\pi n}n^n e^{-n}$). Replacing $m!$, $n!$, and $(n-m)!$ by their approximations and after simplification, we obtain

$$P(X=m) \cong \left(\frac{np}{m}\right)^m \left(\frac{nq}{n-m}\right)^{n-m} \sqrt{\frac{n}{2\pi m(n-m)}}. \tag{14.29}$$

The binomial distribution has the mean value np (see Eq. (14.20). Now let δ denote the deviation of m from np; that is, $\delta = m - np$. Then $n - m = nq - \delta$, and Eq. (14.29) becomes

$$P(X=m) = \frac{1}{\sqrt{2\pi npq(1+\delta/np)(1-\delta/np)}}\left(1+\frac{\delta}{np}\right)^{-(np+\delta)}\left(1-\frac{\delta}{nq}\right)^{-(nq=\delta)}$$

or

$$P(X=m)A = \left(1+\frac{\delta}{np}\right)^{-(np+\delta)}\left(1-\frac{\delta}{nq}\right)^{-(nq-\delta)},$$

where

$$A = \sqrt{2\pi npq\left(1+\frac{\delta}{np}\right)\left(1-\frac{\delta}{nq}\right)}.$$

Then

$$\log(P(X=m)A) \cong -(np+\delta)\log(1+\delta/np) - (nq-\delta)\log(1-\delta/nq).$$

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

Assuming $|\delta| < npq$, so that $|\delta/np| < 1$ and $|\delta/nq| < 1$, this permits us to write the two convergent series

$$\log\left(1+\frac{\delta}{np}\right) = \frac{\delta}{np} - \frac{\delta^2}{2n^2p^2} + \frac{\delta^3}{3n^3p^3} - \cdots,$$

$$\log\left(1-\frac{\delta}{nq}\right) = -\frac{\delta}{nq} - \frac{\delta^2}{2n^2q^2} - \frac{\delta^3}{3n^3q^3} - \cdots.$$

Hence

$$\log(P(X=m)A) \cong -\frac{\delta^2}{2npq} - \frac{\delta^3(p^2-q^2)}{2\times 3n^2p^2q^2} - \frac{\delta^4(p^3+q^3)}{3\times 4n^3p^3q^3} - \cdots.$$

Now, if $|\delta|$ is so small in comparison with npq that we ignore all but the first term on the right hand side of this expansion and A can be replaced by $(2\pi npq)^{1/2}$, then we get the approximation formula

$$P(X=m) = \frac{1}{\sqrt{2\pi npq}} e^{-\delta^2/2npq}. \qquad (14.30)$$

When $\sigma = \sqrt{npq}$, Eq. (14.30) becomes

$$f(m) = P(X=m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\delta^2/2\sigma^2}. \qquad (14.31)$$

This is called the Guassian, or normal, distribution. It is a very good approximation even for quite small values of n.

The Gaussian distribution is a symmetrical bell-shaped distribution about its mean μ, and σ is a measure of the width of the distribution. Fig. 14.4 gives a comparison of the binomial distribution and the Gaussian approximation.

The Gaussian distribution also has a significance far beyond its connection with the binomial distribution. It can be derived mathematically from elementary considerations, and is found to agree empirically with random errors that actually

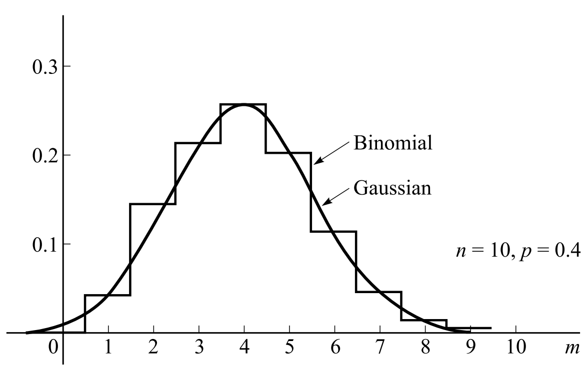

Figure 14.4.

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

occur in experiments. Everyone believes in the Gaussian distribution: mathematicians think that physicists have verified it experimentally and physicists think that mathematicians have proved it theoretically.

One of the main uses of the Gaussian distribution is to compute the probability

$$\sum_{m=m_1}^{m_2} f(m)$$

that the number of successes is between the given limits m_1 and m_2. Eq. (14.31) shows that the above sum may be approximated by a sum

$$\sum \frac{1}{\sqrt{2\pi}\sigma} e^{-\delta^2/2\sigma^2} \tag{14.32}$$

over appropriate values of δ. Since $\delta = m - np$, the difference between successive values of δ is 1, and hence if we let $z = \delta/\sigma$, the difference between successive values of z is $\Delta z = 1/\sigma$. Thus Eq. (14.32) becomes the sum over z,

$$\sum \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \Delta z. \tag{14.33}$$

As $\Delta z \to 0$, the expression (14.33) approaches an integral, which may be evaluated in terms of the function

$$\Phi(z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-z^2/2} dz. \tag{14.34}$$

The function $(\Phi(z)$ is related to the extensively tabulated error function, erf(z):

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-z^2} dz, \quad \text{and} \quad \Phi(z) = \frac{1}{2} \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right).$$

These considerations lead to the following important theorem, which we state without proof: If m is the number of successes in n independent trials with constant probability p, the probability of the inequality

$$z_1 \le \frac{m - np}{\sqrt{npq}} \le z_2 \tag{14.35}$$

approaches the limit

$$\frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2} dz = \Phi(z_2) - \Phi(z_1) \tag{14.36}$$

as $n \to \infty$. This theorem is known as Laplace–de Moivre limit theorem.

To illustrate the use of the result (14.36), let us consider the simple example of a die tossed 600 times, and ask what the probability is that the number of ones will

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

be between 80 and 110. Now $n = 600$, $p = 1/6$, $q = 1 - p = 5/6$, and m varies from 80 to 110. Hence

$$z_1 = \frac{80 - 100}{\sqrt{100(5/6)}} = -2.19 \qquad \text{and} \qquad z_1 = \frac{110 - 100}{\sqrt{100(5/6)}} = 1.09.$$

The tabulated error function gives

$$\Phi(z_2) = \Phi(1.09) = 0.362,$$

and

$$\Phi(z_1) = \Phi(-2.19) = -\Phi(2.19) = -0.486,$$

where we have made use of the fact that $\Phi(-z) = -\Phi(z)$, you can check this with Eq. (14.34). So the required probability is approximately given by

$$0.362 - (-0.486) = 0.848.$$

Continuous distributions

So far we have discussed several discrete probability distributions: since measurements are generally made only to a certain number of significant figures, the variables that arise as the result of an experiment are discrete. However, discrete variables can be approximated by continuous ones within the experimental error. Also, in some applications a discrete random variable is inappropriate. We now give a brief discussion of continuous variables that will be denoted by x. We shall see that continuous variables are easier to handle analytically.

Suppose we want to choose a point randomly on the interval $0 \le x \le 1$, how shall we measure the probabilities associated with that event? Let us divide this interval $(0, 1)$ into a number of subintervals, each of length $\Delta x = 0.1$ (Fig. 14.5), the point x is then equally likely to be in any of these subintervals. The probability that $0.3 < x < 0.6$, for example, is 0.3, as there are three favorable cases. The probability that $0.32 < x < 0.64$ is found to be $0.64 - 0.32 = 0.32$ when the interval is divided into 100 parts, and so on. From these we see that the probability for x to be in a given subinterval of (0, 1) is the length of that subinterval. Thus

$$P(a < x < b) = b - a, \qquad 0 \le a \le b \le 1. \tag{14.37}$$

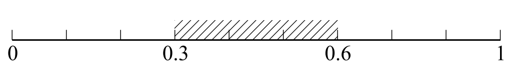

Figure 14.5.

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

The variable x is said to be uniformly distributed on the interval $0 \le x \le 1$. Expression (14.37) can be rewritten as

$$P(a < x < b) = \int_a^b dx = \int_a^b 1dx.$$

For a continuous variable it is customary to speak of the probability density, which in the above case is unity. More generally, a variable may be distributed with an arbitrary density $f(x)$. Then the expression

$$f(z)dz$$

measures approximately the probability that x is on the interval

$$z < x < z + dz.$$

And the probability that x is on a given interval (a, b) is

$$P(a < x < b) = \int_a^b f(x)dx \tag{14.38}$$

as shown in Fig. 14.6.

The function $f(x)$ is called the probability density function and has the properties:

(1) $f(x) \ge 0, \qquad (-\infty < x < \infty)$;

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$, a real-valued random variable must lie between $\pm\infty$.

The function

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du \tag{14.39}$$

defines the probability that the continuous random variable X is in the interval $(-\infty, x)$, and is called the cumulative distributive function. If $f(x)$ is continuous, then Eq. (14.39) gives

$$F'(x) = f(x)$$

and we may speak of a probability differential $dF(x) = f(x)dx$.

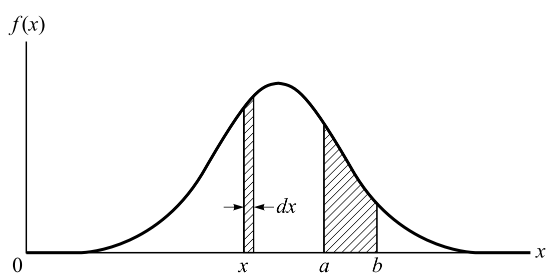

Figure 14.6.

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press

By analogy with those for discrete random variables the expected value or mean and the variance of a continuous random variable X with probability density function $f(x)$ are defined, respectively, to be:

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx; \tag{14.40}$$

$$\text{Var}(X) = \sigma^2 = E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx. \tag{14.41}$$

The Gaussian (or normal) distribution

One of the most important examples of a continuous probability distribution is the Gaussian (or normal) distribution. The density function for this distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty, \tag{14.42}$$

where μ and σ are the mean and standard deviation, respectively. The corresponding distribution function is

$$F(x) = P(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(u-\mu)^2/2\sigma^2} du. \tag{14.43}$$

The standard normal distribution has mean zero ($\mu = 0$) and standard deviation ($\sigma = 1$)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{14.44}$$

Any normal distribution can be 'standardized' by considering the substitution $z = (x-\mu)/\sigma$ in Eqs. (14.42) and (14.43). A graph of the density function (14.44), known as the standard normal curve, is shown in Fig. 14.7. We have also indicated the areas within 1, 2 and 3 standard deviations of the mean (that is between $z = -1$ and $+1$, -2 and $+2$, -3 and $+3$):

$$P(-1 \le Z \le 1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-z^2/2} dz = 0.6827;$$

$$P(-2 \le Z \le 2) = \frac{1}{\sqrt{2\pi}} \int_{-2}^{2} e^{-z^2/2} dz = 0.9545;$$

$$P(-3 \le Z \le 3) = \frac{1}{\sqrt{2\pi}} \int_{-3}^{3} e^{-z^2/2} dz = 0.9973.$$

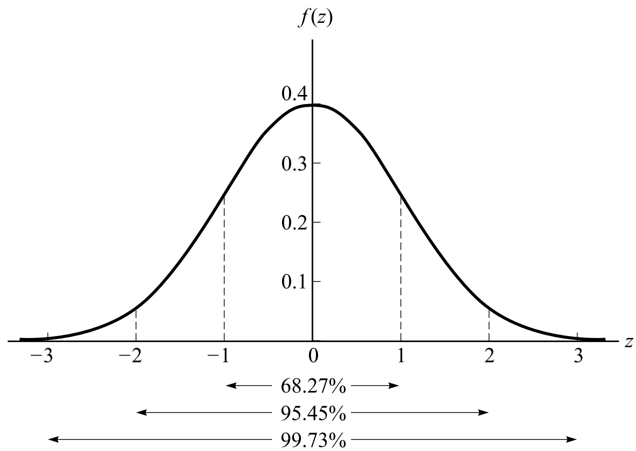

Figure 14.7.

The above three definite integrals can be evaluated by making numerical approximations. A short table of the values of the integral

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2} dt \left(= \frac{1}{2} \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{-t^2} dt \right)$$

is included in Appendix 3. A more complete table can be found in *Tables of Normal Probability Functions*, National Bureau of Standards, Washington, DC, 1953.

The Maxwell–Boltzmann distribution

Another continuous distribution that is very important in physics is the Maxwell–Boltzmann distribution

$$f(x) = 4a\sqrt{\frac{a}{\pi}} x^2 e^{-ax^2}, \quad 0 \le x < \infty, \quad a > 0, \tag{14.45}$$

where $a = m/2kT$, m is the mass, T is the temperature (K), k is the Boltzmann constant, and x is the speed of a gas molecule.

Problems

14.1 If a pair of dice is rolled what is the probability that a total of 8 shows?

14.2 Four coins are tossed, and we are interested in the number of heads. What is the probability that there is an odd number of heads? What is the probability that the third coin will land heads?

14.3 Two coins are tossed. A reliable witness tells us 'at least 1 coin showed heads.' What effect does this have on the uniform sample space?

14.4 The tossing of two coins can be described by the following sample space:

Event	no heads	one head	two head
Probability	1/4	1/2	1/4

What happens to this sample space if we know at least one coin showed heads but have no other specific information?

14.5 Two dice are rolled. What are the elements of the sample space? What is the probability that a total of 8 shows? What is the probability that at least one 5 shows?

14.6 A vessel contains 30 black balls and 20 white balls. Find the probability of drawing a white ball and a black ball in succession from the vessel.

14.7 Find the number of different arrangements or permutations consisting of three letters each which can be formed from the seven letters A, B, C, D, E, F, G.

14.8 It is required to sit five boys and four girls in a row so that the girls occupy the even seats. How many such arrangements are possible?

14.9 A balanced coin is tossed five times. What is the probability of obtaining three heads and two tails?

14.10 How many different five-card hands can be dealt from a shuffled deck of 52 cards? What is the probability that a hand dealt at random consists of five spades?

14.11 (*a*) Find the constant term in the expansion of $(x^2 + 1/x)^{12}$.
(*b*) Evaluate 50!.

14.12 A box contains six apples of which two are spoiled. Apples are selected at random without replacement until a spoiled one is found. Find the probability distribution of the number of apples drawn from the box, and present this distribution graphically.

14.13 A fair coin is tossed six times. What is the probability of getting exactly two heads?

14.14 Suppose three dice are rolled simultaneously. What is the probability that two 5*s* appear with the third face showing a different number?

14.15 Verify that $\sum_{m=0}^{\infty} P(X = m) = 1$ for the Poisson distribution.

14.16 Certain processors are known to have a failure rate of 1.2%. There are shipped in batches of 150. What is the probability that a batch has exactly one defective processor? What is the probability that it has two?

14.17 A Geiger counter is used to count the arrival of radioactive particles. Find:
(*a*) the probability that in time t no particles will be counted;
(*b*) the probability of exactly one count in time t.

14.18 Given the density function $f(x)$

$$f(x) = \begin{cases} kx^2 & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases} :$$

(*a*) find the constant k;
(*b*) compute $P(1 < x < 2)$;
(*c*) find the distribution function and use it to find $P(1 < x \leq (2))$.

https://doi.org/10.1017/CBO9780511755781.015 Published online by Cambridge University Press