3
Random Variables

This chapter extends the ideas of mean, standard deviation, and normal approximation to distributions more general than the binomial. This involves sums and averages of randomly produced numbers. Random variables, introduced in Section 3.1, provide a good notation for this purpose. The concept of the expectation or mean of a random variable is the subject of Section 3.2. Then standard deviation and the normal approximation appear in Section 3.3. In these first three sections, attention is restricted to random variables with a finite number of possible values. The ideas are extended to random variables with an infinite sequence of possible values in Section 3.4, then to random variables with a continuous distribution in the following chapters.

3.1 Introduction

The number of heads in four tosses of a coin could be any one of the possible values $0, 1, 2, 3, 4$. The term *random variable* is now introduced for something like the number of heads, which might be one of several possible values, with a distribution of probabilities over this set of values. Typically, capital letters X, Y, Z, etc., are used to denote random variables. For example, X might stand for "the number obtained by rolling a die", Y for "the number of heads in four coin tosses", and Z for "the suit of a card dealt from a well-shuffled deck". This is not really a new idea, rather a compact notation for the familiar idea of something or other picked at random according to a probability distribution.

J. Pitman, *Probability*

The *range* of a random variable X is the set of all *possible values* that X might produce. This section only considers random variables with a finite range. But infinite ranges will appear in later sections. Usually, and unless otherwise specified in the following development, the range of a random variable is assumed to be a set of numbers. In case not, the nature of the range can be indicated by a change in terminology. For example, *random pair*, *random sequence*, or *random permutation*. In the following table, Z might be called a *random suit*.

TABLE 1. Some random variables and their ranges.

Random variable	Description	Range
X	Number on a die	$\{1,2,3,4,5,6\}$
Y	Number of heads in 4 coin tosses	$\{0,1,2,3,4\}$
Z	Suit of a card	$\{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$

Distribution of X

A statement about a random variable, such as "$X \leq 3$", defines an event. The event occurs if the statement is true, and does not occur if the statement is false.

TABLE 2. Some events determined by X, the number on a die.

Verbal description of event	Notation	Subset of range	Probability
1. Number on the die is less than or equal to 3	$X \leq 3$	$\{1,2,3\}$	$1/2$
2. Number on the die is 6	$X = 6$	$\{6\}$	$1/6$
3. Number on the die is less than or equal to x	$X \leq x$	$\{1,2,\ldots,x\}$	$x/6$
4. Number on the die is x	$X = x$	$\{x\}$	$1/6$
5. Number on the die is in the subset B	$X \in B$	B	$\#(B)/6$

In lines 3 and 4 of the table, x denotes an arbitrary element of the range of X. In line 5, B is a generic subset of the range of X. Events defined by statements about a random variable X are called *events determined by* X. Every such event can be written as "$X \in B$" where B is the set of possible values of X for which the statement is true. The probability of this event is written $P(X \in B)$, or simply $P(B)$. The notation $P(B)$ is familiar as the probability of getting a value in B. The notation $P(X \in B)$ shows this probability refers to the random variable X. As B varies over subsets of the range of X, these probabilities must form a distribution, called *the distribution of* X. Assuming that X has only a finite number of possible

values, the distribution of X is determined by the probabilities of individual values,

$$P(X = x) \qquad x \in \text{range of } X$$

via the addition rule

$$P(X \in B) = \sum_{x \in B} P(X = x)$$

Here it is assumed that the random variable X has a uniquely specified value, no matter what happens. So the events $(X = x)$ as x varies over the range of X are mutually exclusive and exhaustive, and their probabilities must add up to 1. By similar reasoning, $P(X \in B)$ is obtained by summing just over those values x in B. The probabilities $P(X = x)$ can be displayed in a distribution table or histogram, or given by a formula.

Dummy variables. There is nothing sacred about the use of the symbol x as a generic possible value of X. You could just as well use k or i or any other lowercase letter. For example, if X is the number of heads in four coin tosses, it makes perfect sense to write both

$$P(X = k) = \binom{4}{k} 2^{-4} \qquad (k = 0, \ldots, 4)$$

$$P(X \leq 2) = \sum_{i=0}^{2} \binom{4}{i} 2^{-4}$$

Here k and i are called *dummy variables*. It is a useful convention to reserve capital letters for random variables, small letters for dummy variables. Often a matching lowercase letter is used to denote a generic possible value for an uppercase random variable. But this is not always convenient. So be prepared for statements like

$$P(X = v) = P(Y = v)$$

which means that X and Y have the same chance of being equal to v.

Functions

Often a random variable of interest, X say, is expressed as a function of another random variable W:

$$X = g(W)$$

Here g is a function defined on the range of W with values in the range of X. Such a function is a deterministic rule. The rule is that if W has value w, then X has value $g(w)$, uniquely determined by w, for every possible value w of W. Put another way,

X gives a less detailed description of what is happening than W. The distribution of X can be derived from that of W, because any event defined by X can be written in terms of $g(W)$ and hence in terms of W. As the next example shows, this is just a new way to say something familiar.

Example 1. Number of heads.

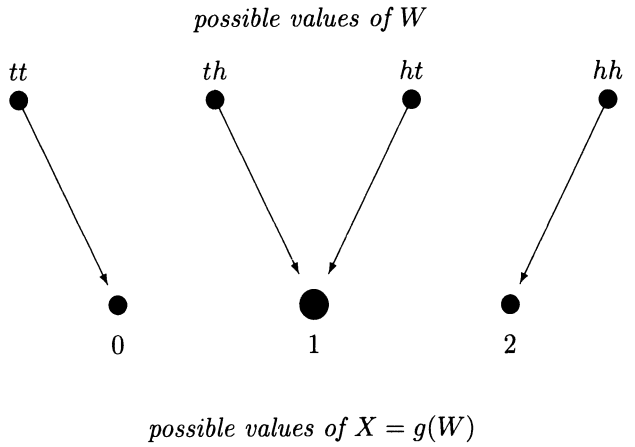
Let X be the number of heads in two tosses of a fair coin. The distribution of X is the binomial distribution with parameters $n = 2$ and $p = 1/2$, as discussed in Section 2.1:

x	0	1	2
$P(X = x)$	1/4	1/2	1/4

The probabilities of $1/4$, $1/2$, and $1/4$ were obtained from the natural outcome space for two coin tosses: $\{hh, ht, th, tt\}$, by assuming the two tosses were independent. Let W represent which of these outcomes appeared. Once the random outcome W of both tosses becomes known, the number of heads X is completely determined by $X = g(W)$ where g is the function defined by the following table:

Outcome of tosses w	tt	th	ht	hh
Number of heads $g(w)$	0	1	1	2

The same relationship is displayed in the following diagram:



As the blobs and arrows suggest, the probability of each possible value x of X is the sum of the probabilities of those w for which $g(w) = x$. For $x = 2$ and 0 there is a unique w giving $g(w) = x$, so $P(X = x) = 1/4$ for these x. But there are two outcomes w giving $g(w) = 1$, so $P(X = 1) = 1/4 + 1/4 = 1/2$.

The distribution of $X = g(W)$. As in the last example with two trials, the number of successes in n trials can be regarded as a function of the detailed sequential outcome of all trials. To get the probability of a particular number k of successes, add the probabilities of all sequences giving rise to k successes. The same method gives a general formula for the distribution of $X = g(W)$ in terms of the distribution of W. Keep in mind that while a function g must assign to each w a unique value of x, many values of w may be assigned the same x. The event $(X = x)$ is the event that W has a value w such that $g(w) = x$. By the addition rule for probability, $P(X = x)$ is the sum of the probabilities $P(W = w)$ over all w such that $g(w) = x$:

$$P(X = x) = P(g(W) = x) = \sum_{w:g(w)=x} P(W = w)$$

Given a random variable X, new random variables are created by common numerical functions, for example,

$$2X \qquad 3X - 5 \qquad X^2 \qquad |X - 2|$$

To illustrate, if the value of X turns out to be -3, the values of these four variables are

$$-6 \qquad -14 \qquad 9 \qquad 5$$

Assuming the distribution of X is known, the probability of an event determined by a function of X is often found most simply by manipulating the statement of the event. The result of the manipulation is that the event in question occurs precisely when X falls in some set of values. To illustrate, suppose X has uniform distribution on the 19 integers $\{-9, -8, \ldots, 8, 9\}$. Then

$$\begin{aligned} P(2X \le 5) &= P(X \le 5/2) = 12/19 \\ P(3X - 5 \le 5) &= P(X \le 10/3) = 13/19 \\ P(X^2 \le 5) &= P(-\sqrt{5} \le X \le \sqrt{5}) = 5/19 \\ P(|X - 2| \le 5) &= P(-5 \le X - 2 \le 5) \\ &= P(-3 \le X \le 7) = 11/19 \end{aligned}$$

Events like the last one turn up in prediction problems. If you try predicting the value of X by guessing that X is 2, then $|X - 2|$ is how far off your prediction is. And $P(|X - 2| \le 5)$, found above, is the chance that your prediction is off by 5 or less.

Technical remark. In a more mathematical development of these ideas, it is necessary to say precisely what kind of mathematical object is a random variable. In the usual treatment, a random variable X is, by definition, a numerical function $X(w)$ defined on some basic space of possible outcomes w, where a probability

distribution is given. For example, X representing a number of heads as in Example 1 would be the function $X(w)$, denoted $g(w)$ in that example, giving the number of heads as a function of a more complete description of the outcome. Then $P(X \in B) = P(\{w : X(w) \in B\})$ defines the distribution of X in terms of probability on the basic outcome space. With this formalism, a function h defined on the range of X defines another random variable $h(X)$, the composition of h and X, which is the function whose value for outcome w is $h(X(w))$.

Joint Distributions

Given two random variables X and Y defined in the same setting, we can consider their *combined* or *joint* outcome (X, Y) as a random pair of values. By definition, (X, Y) has value (x, y) if X has value x and Y has value y. Thus the event that $((X, Y) = (x, y))$ is the intersection of the events $(X = x)$ and $(Y = y)$, and is usually denoted $(X = x, Y = y)$. So commas mean intersections in statements about random variables.

The range of the joint outcome (X, Y) is the set of all ordered pairs (x, y) with x in the range of X, y in the range of Y, and $P(X = x, Y = y) > 0$. If the range of X is represented by points on a horizontal line, and the range of Y by points on a vertical line, then the range of (X, Y) is represented by a set of points in the plane. Alternatively, the range of (X, Y) may be represented by a set of paths through a tree diagram, as in Chapter 1.

The distribution of (X, Y) is called the *joint distribution* of X and Y. This distribution is determined by the probabilities

$$P(x, y) = P(X = x, Y = y)$$

which must satisfy

$$P(x, y) \geq 0 \qquad \text{and} \qquad \sum_{\text{all } (x,y)} P(x, y) = 1$$

Example 2. Two draws at random without replacement.

Let X and Y be the first and second draws made at random without replacement from a box containing three tickets numbered 1, 2, and 3. Assuming all six possible pairs of draws are equally likely, the joint distribution of X and Y is displayed as follows. The entry at position (x, y) is $P(x, y) = P(X = x, Y = y)$, the chance that the first draw is x and the second is y. Contrary to convention for matrices, here the first index x is for columns, increasing from left to right, and the second index y is for rows, increasing from bottom to top. This is to make the table consistent with conventional (x, y) co-ordinates in the plane, as in Figure 1 on page 148.

TABLE 3. Joint distribution table for (X, Y)

		possible values for X			distn. of Y
		1	2	3	(row sums)
possible	3	1/6	1/6	0	1/3
values	2	1/6	0	1/6	1/3
for Y	1	0	1/6	1/6	1/3
	distn. of X (column sums)	1/3	1/3	1/3	1 (total sum)

As in this example, the distribution of X can be obtained using the following:

Marginal Probabilities

$$P(X = x) = \sum_{\text{all } y} P(x, y)$$

where the sum is over all possible y in the range of Y.

This is just the basic addition rule for probabilities, since the events $(X = x, Y = y)$ form a partition of $(X = x)$ as y varies over the range of Y. The sum is over all entries in column x of the distribution table. These sums can be displayed as above to show the distribution table for X in a row along the bottom margin of the table. Similarly, the distribution of Y defined by

$$P(Y = y) = \sum_{\text{all } x} P(x, y)$$

can be displayed in a column on the right margin of the table. For this reason, when a joint distribution of X and Y is considered, the distribution of X and the distribution of Y are often called *marginal distributions.*

Same random variable or same distribution? In the last example, while the two random variables X and Y have *identical distributions*, it would be wrong to say they were equal. Indeed, for the two draws without replacement,

$$P(X = Y) = 0$$

so X is certainly not equal to Y. A second example: if X is the number of heads in ten tosses of a fair coin, and Y is the number of tails in those ten tosses, then X

and Y have identical distributions. Still, X and Y are not equal, since, for instance, $X = 6$ makes $Y = 4$. However X equals $10 - Y$, because no matter what the pattern of heads and tails, the number of heads is 10 minus the number of tails. That is to say X is certain to equal $10 - Y$, or $P(X = 10 - Y) = 1$. The next box summarizes this distinction.

Random Variables with the Same Distribution

Random variables X and Y have the *same* or *identical distribution* if X and Y have the same range, and for every value v in this range,

$$P(X = v) = P(Y = v).$$

Change of Variable Principle

If X has the same distribution as Y, then any statement about X has the same probability as the corresponding statement about Y, and $g(X)$ has the same distribution as $g(Y)$, for any function g. For example,

$$P(a \le X \le b) = P(a \le Y \le b) \text{ for all } a \text{ and } b,$$

and X^2 has the same distribution as Y^2.

Equality of Random Variables

Random variables X and Y are *equal*, written $X = Y$, if $P(X = Y) = 1$. In particular, if no matter what the outcome, the value of X equals the value of Y, then $X = Y$.

If two random variables are equal, then they have the same distribution. But random variables with the same distribution need not be equal.

The change of variable principle is an immediate consequence of the definition of equality in distribution. A later subsection on symmetry shows how the change of variable principle can be used to avoid unnecessary calculations.

Technical remark. The definition of equality of X and Y allows X and Y to differ on some exceptional set of outcomes that is assigned probability zero. This flexibility in the definition is of little significance for random variables with a finite range, but is convenient for random variables with infinite range, considered in later sections.

Computing probabilities from a joint distribution. Once the joint distribution of X and Y has been calculated, the probability of any event defined in terms of X and Y can be found. Simply sum the probabilities $P(x, y)$ over the relevant set of pairs (x, y):

Probabilities of Events Determined by X and Y

The probability that X and Y satisfy some condition is the sum of $P(x,y)$ over all pairs (x,y) satisfying that condition. For instance

$$P(X<Y) = \sum_{(x,y):x<y} P(x,y) = \sum_{\text{all } x} \sum_{y:y>x} P(x,y)$$

$$P(X=Y) = \sum_{(x,y):x=y} P(x,y) = \sum_{\text{all } x} P(x,x)$$

Distribution of a function of X and Y. The distribution of any function of (X,Y), for example

$$X+Y \qquad X-Y \qquad XY \qquad \min(X,Y) \qquad \max(X,Y)$$

can be obtained from the joint distribution of X and Y. For example,

$$P(X+Y=z) = \sum_{(x,y):x+y=z} P(x,y) = \sum_{\text{all } x} P(x,z-x).$$

There is a similar formula for any function $g(X,Y)$: the probability $P[g(X,Y)=z]$ is the sum of $P(x,y)$ over all pairs (x,y) with $g(x,y)=z$.

Example 3. Sum of the draws.

Problem. Calculate the distribution of $X+Y$ for two random draws X and Y from a box containing $\{1,2,3\}$: (a) without replacement, (b) with replacement.

Solution. (a) From the joint distribution table given earlier for draws without replacement, the possible values of the sum $S=X+Y$ are 3, 4, and 5. By inspection of the table, each possible value s for S corresponds to exactly two possible pairs (x,y), each with probability $\frac{1}{6}$. Hence the distribution of S is given by the following table:

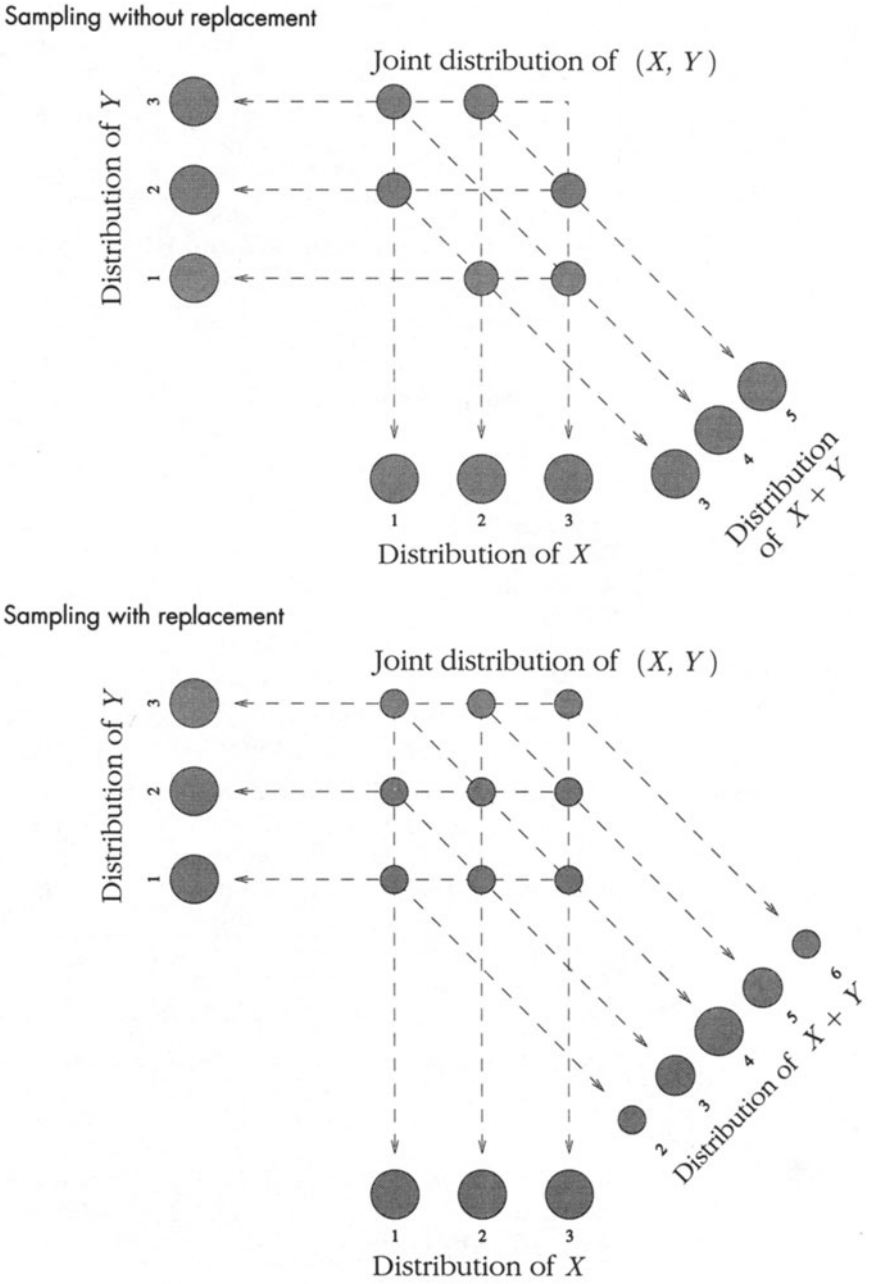
s	3	4	5
$P(S=s)$	1/3	1/3	1/3

(b) If the draws are made with replacement, then the joint probabilities are

$$P(x,y) = 1/9 \qquad (1 \le x \le 3, \quad 1 \le y \le 3)$$

Now, there is one possible pair adding to 2, two possible pairs adding to 3, three adding to 4, two to 5, and one to 6. Thus for draws with replacement the distribution of S is given by the table:

FIGURE 1. Distributions for sampling with and without replacement from $\{1, 2, 3\}$. Refer to Example 3. In each case the joint distribution of (X, Y) is represented by a pattern of blobs, with the area of the blob over (x, y) proportional to $P(x, y)$. The distributions of X, Y, and $X + Y$ are displayed similarly around the edges of the joint pattern. Probabilities in these distributions are obtained by adding probabilities from the joint distribution as indicated by the arrows.



s	2	3	4	5	6
$P(S = s)$	1/9	2/9	3/9	2/9	1/9

These calculations are illustrated in Figure 1.

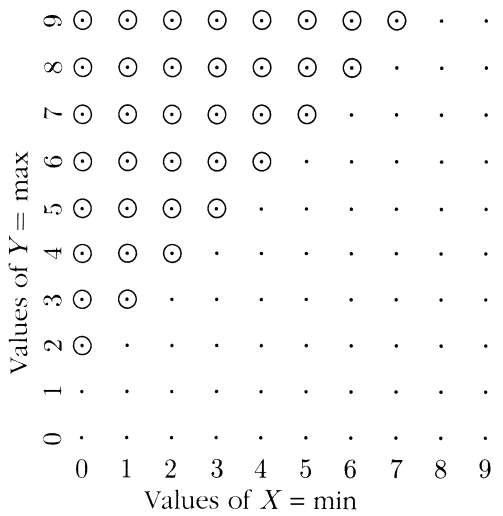
Discussion. The example shows that to find the distribution of a function of two random variables X and Y, such as their sum, you must think in terms of the joint distribution of X and Y. Knowing that both X and Y are uniform on $\{1, 2, 3\}$ is not enough to determine the distribution of $X + Y$.

Example 4. Minimum and maximum.

Let X be the minimum and Y the maximum of three digits picked at random without replacement from $\{0, 1, \ldots, 9\}$.

Problem 1. Find the joint distribution of X and Y.

Solution. Because the sampling is done without replacement the three digits drawn must be distinct, so the only possible pairs of values for the minimum and maximum are integers x and y with



$$0 \le x \le 7, \quad x + 2 \le y, \quad y \le 9,$$

as marked in the diagram on the left. To find the chance of such a pair, forget the order in which the digits come, and think about what subset of three digits is chosen from $\{0, 1, \ldots, 9\}$. Every subset has the same chance of being chosen, 1 in $\binom{10}{3}$, or 1/120. To illustrate, for a minimum of 4 and a maximum of 7 there are just 2 possible subsets, $\{4, 5, 7\}$ and $\{4, 6, 7\}$. In this case, there are $7 - 4 - 1 = 2$ ways to pick the intermediate number. So

$$P(X = 4, Y = 7) = (7 - 4 - 1)/120$$

In general, for a minimum of x and a maximum of y, there are $y - x - 1$ possible subsets, one for each possible value of the third number between x and y. Therefore, for possible pairs x and y as above

$$P(X = x, Y = y) = (y - x - 1)/120$$

Problem 2. Find the distribution of $Z = Y - X$, the maximum minus the minimum.

Solution. The possible values of Z are clearly $2, 3, \ldots, 9$. Any one of these possible values, z say, must come from a (min, max) pair (x, y) with a difference of $y - x = z$. Every such pair (x, y) has the same probability

$$(y - x - 1)/120 = (z - 1)/120$$

For $z = 9$ there is one pair $(0, 9)$, for $z = 8$ there are 2 pairs $(0, 8)$ and $(1, 9)$, and so on. In general, there are $10 - z$ possible pairs (x, y) with $y - x = z$. Therefore,

$$P(Z = z) = (10 - z)(z - 1)/120 \qquad (z = 2, 3, \ldots, 9)$$

To check, the sum of these probabilities from $z = 2$ to 9 is

$$((8\times1) + (7\times2) + (6\times3) + (5\times4) + (4\times5) + (3\times6) + (2\times7) + (1\times8))\ /120 = 1$$

Conditional Distributions

The basic rules of probability imply that for any given event A, and any random variable Y, the collection of conditional probabilities

$$P(Y \in B \mid A) = \frac{P[(Y \in B) \text{ and } A]}{P(A)}$$

defines a probability distribution as B varies over subsets of the range Y. This distribution is called the *conditional distribution of Y given A*. Intuitively, this is the appropriate revision of the distribution of Y given the information that event A has occurred. For Y with a finite range the conditional distribution of Y given A is specified by the conditional probabilities

$$P(Y = y \mid A) \quad \text{for } y \in \text{range of } Y.$$

The rules of a probability distribution imply $P(Y \in B \mid A) = \sum_{y \in B} P(Y = y \mid A)$. Most often the conditional distribution of Y given A is considered for each A of the form $(X = x)$ for some random variable X.

Conditional Distribution of Y Given X = x

For each possible value x of X, as y varies over the range of Y the probabilities $P(Y = y \mid X = x)$ define a probability distribution over the range of Y. This probability distribution, which may depend on the given value x of X, is called the *conditional distribution of Y given $X = x$*.

The given value x of X can be thought of as a *parameter* in the distribution of Y given $X = x$. If the joint distribution of X and Y is tabulated, then for given x the conditional probabilities $P(Y = y \mid X = x)$, are found from the joint distribution table by lifting out column x of the table and renormalizing the probabilities in this column by their sum, which is $P(X = x)$. Similarly, for given y, the probabilities $P(X = x \mid Y = y)$ for x in the range of X are found by lifting out row y from the table of joint probabilities and renormalizing this row of probabilities by their sum, which is $P(Y = y)$.

If the marginal (*unconditional*) distribution of X is known, together with the conditional distribution of Y given $X = x$ for all possible values x of X, the joint distribution of X and Y is found using the

Multiplication Rule

$$P(X = x, Y = y) = P(X = x)P(Y = y \mid X = x)$$

In this section conditional distributions serve only to motivate the following definition of independent random variables. See Section 6.1 (which can be read immediately) for a detailed discussion of conditional distributions for dependent random variables.

Independence

Intuitively, random variables X and Y are independent when the probabilities for various values of Y are unaffected by conditioning on the value of X. This is just a restatement in terms of random variables of the relation of independence between draws, trials, etc., as discussed in Chapter 1. For calculations with independent random variables, the simplest definition of independence is the following one using the product rule:

Independent Random Variables

Random variables X and Y are *independent* if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \text{for all } x \text{ and } y$$

If X and Y are independent random variables, then every event determined by X is independent of every event determined by Y:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Conceptually, independence means that conditioning on a given value of X does not affect the distribution of Y, and vice-versa. Thus the above definition of independence can be re-expressed as follows in terms of conditional distributions:

Conditional Distributions and Independence

The following three conditions are equivalent:

- X and Y are independent;
- the conditional distribution of Y given $X = x$ does not depend on x;
- the conditional distribution of X given $Y = y$ does not depend on y.

Example 5. Independent or not?

Problem. A box of 10 tickets contains some number r of red tickets. The rest are green. A sample of 100 tickets is drawn at random with replacement. Then a second sample of 100 tickets is drawn at random with replacement. Let X_1 be the number of red tickets in the first sample, and X_2 the number in the second sample. Are X_1 and X_2 independent?

Solution 1. If you regard r as known, then no matter how many red tickets you see in the first 100 draws, the second 100 draws is still a random sample with replacement from r red and $10 - r$ green tickets. Thus X_1 and X_2 are independent random variables, each with binomial distribution with parameters $n = 100$ and $p = r/10$.

Solution 2. On the other hand, if you don't know r, it seems intuitively obvious that X_1 and X_2 are dependent. For if you saw 53 reds in the first 100 draws, you would be inclined to guess there were around 5 red tickets in the box, and expect to see around 50% red on the next 100 draws. Whereas if you saw 17 reds in the first 100 draws, you would guess that 2 of the 10 tickets were red, and expect to see only 20% or so red on the next 100 draws. Thus, knowing the value of X_1 affects the chances of events determined by X_2, so X_1 and X_2 are dependent.

Discussion. Which solution is correct? It depends on whether r is regarded as a known constant, as in Solution 1, or the value of a random variable, R say, as in Solution 2. Solution 2 can be made more precise by assuming that *conditionally* on the event $(R = r)$ the random variables X_1 and X_2 are independent, with binomial $(100, r/10)$ distribution, just as if r were known as in Solution 1. But unconditionally these variables will be dependent, for the reasons given in Solution 2. Does it make sense to think of r as the value of a random variable R? With a frequency interpretation of probability, it makes sense only if the way the composition of the box was determined is regarded as somehow repeatable. The probabilities $P(R = r)$ for $0 \le r \le 10$ would then be long-run frequencies of different compositions. With a subjective interpretation of

probability, $P(R = r)$ might be assigned according to your own opinion about the unknown number of reds in the box, even if there is no notion of repetitions.

Several Random Variables

The joint distribution of several random variables $X_1, X_2, \ldots, X_n$ is defined just as for two random variables by the *joint probabilities*

$$P(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

for all possible values x_i of each X_i. Note that the commas signify an *intersection* of events. So $P(x_1, \ldots, x_n)$ is the probability that X_i has value x_i for every $1 \le i \le n$. This concept will now be illustrated by a number of examples.

Random permutations. A permutation of $\{1, 2, \ldots, n\}$ is a sequential ordering of the n numbers with no repeats. A *random permutation* of $\{1, 2, \ldots, n\}$ is a permutation picked uniformly at random from all $n!$ possible permutations of $\{1, 2, \ldots, n\}$. There are many ways to generate a random permutation. For example,

- Suppose tickets numbered $1, 2, \ldots, n$ are placed in a box and drawn one by one at random without replacement. Let X_i be the number of the ith ticket drawn, $1 \le i \le n$. Then $(X_1, X_2, \ldots, X_n)$ is a random permutation of $\{1, 2, \ldots, n\}$.
- Suppose cards numbered $1, 2, \ldots, n$ are thoroughly shuffled. Let Y_i be the number of the ith card from the top of the deck. Then $(Y_1, Y_2, \ldots, Y_n)$ is a random permutation of $\{1, 2, \ldots, n\}$.

Example 6. Joint distribution of a random permutation.

Problem 1. Describe the joint distribution of a random permutation of $\{1, 2, \ldots, n\}$, that is the common joint distribution of $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_n)$.

Solution. Informally the answer is just "the uniform distribution over all $n!$ possible permutations of $\{1, \ldots, n\}$". To illustrate for $n = 3$, (X_1, X_2, X_3) is equally likely to be any one of the 3! = 6 permutations

$$(1,2,3),\ (1,3,2),\ (2,1,3),\ (2,3,1),\ (3,1,2),\ (3,2,1)$$

and so is (Y_1, Y_2, Y_3). To state this in a formula for a general n, the joint probabilities

$$P(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n) = P(Y_1 = x_1, \ldots, Y_n = x_n)$$

are given by

$$P(x_1, \ldots, x_n) = \begin{cases} 1/n! & \text{if } (x_1, \ldots, x_n) \text{ is a permutation of } \{1, 2, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}$$

Discussion. Note that $P(x_1, \ldots, x_n)$ is a symmetric function of $(x_1, \ldots, x_n)$, as defined in Section 3.6, because for any rearrangement of the order of terms in a sequence, the original sequence is a permutation if and only if the rearranged sequence is a permutation. This symmetry property, studied further in Section 3.6, explains the simple solutions of both the next problem and the problem of Example 1.4.7.

Problem 2. For each $1 \leq j \leq n$, find the distribution of X_j for $(X_1, X_2, \ldots, X_n)$ a random permutation of $\{1, 2, \ldots, n\}$.

Solution. For each $1 \leq x \leq n$, the probability $P(X_j = x)$ is the number of permutations with x in the jth place, divided by $n!$. But if value x is fixed in the jth place, the values in the remaining $n - 1$ places can be any permutation of the set $\{1, 2, \ldots, n\}$ with x deleted. Since there are $(n-1)!$ such permutations, whatever $x \in \{1, 2, \ldots, n\}$, $P(X_j = x) = (n-1)!/n! = 1/n$. Conclusion: for every $1 \leq j \leq n$, the distribution of X_j is uniform on $\{1, 2, \ldots, n\}$.

Independence of several variables. Random variables $X_1, \ldots, X_n$ are *independent* if their joint probabilities are products of their marginal probabilities:

$$P(x_1, x_2, \ldots, x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n)$$

for all possible values x_i of each X_i. Summing these probabilities over all $(x_1, \ldots, x_n)$ such that $x_i \in A_i$ shows that events of the form $(X_i \in A_i)$ determined by independent random variables X_i are independent:

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2) \cdots P(X_n \in A_n)$$

Here for each i the set A_i can be any subset of the range of possible values of X_i. The results of the next three paragraphs are consequences of this formula.

Functions of independent random variables are independent. If $X_j, 1 \leq j \leq n$, are independent random variables, then so are the random variables Y_j defined by $Y_j = f_j(X_j)$ for arbitrary functions f_j defined on the range of X_j.

Disjoint blocks of independent random variables are independent. For example, if $X_1, X_2, \ldots, X_6$ are independent, then (X_1, X_2), (X_3, X_4), and (X_5, X_6) are three independent random pairs. These properties can be combined:

Functions of disjoint blocks of independent random variables are independent. For example, if $X_1, \ldots, X_5$ are independent positive random variables, then so are Y_1, Y_2, and Y_3 defined by $Y_1 = 5X_3 + \sqrt{X_5}$, $Y_2 = X_4 X_2$, $Y_3 = X_1$.

Repeated trials. Independent random variables with the same distribution, for example, repeated draws at random with replacement from some population, or repeated rolls of a die (perhaps biased) are called *repeated trials*. Independent trials that result in one of two possible outcomes, say success or failure, with constant probability p of success on each trial, as studied in Chapter 2, are called *Bernoulli*(p) trials. The number of successes S_n in n Bernoulli trials can be represented as

$$S_n = X_1 + X_2 + \cdots + X_n$$

where X_i is the *indicator* of success on trial i, that is to say the random variable that is 1 if trial i is a success and 0 if trial i is a failure. The sum simply counts the number of 1's, that is the number of successes in the n trials. The sequence $X_1, X_2, \ldots, X_n$ is a sequence of n independent random variables, each with the Bernoulli(p) distribution on $\{0, 1\}$ defined at the end of Section 1.3. The Bernoulli(p) distribution of each X_i is the special case $n = 1$ of the binomial (n, p) distribution of the number of successes S_n in n trials, analyzed in Chapter 2. The next two sections show how the representation of S_n as the sum of n independent indicator variables leads to extensions of the law of large numbers and the normal approximation described in Chapter 2 to sums of independent random variables X_i with any common distribution over a finite set of possible values.

Here is the generalization of the binomial distribution that describes the joint distribution of counts in any finite number m of categories in independent trials.

Multinomial Distribution

Let N_i denote the number of results in category i in a sequence of independent trials with probability p_i for a result in the ith category on each trial, $1 \le i \le m$, where $p_1 + \cdots + p_m = 1$. Then for every m-tuple of non-negative integers $(n_1, n_2, \ldots, n_m)$ with sum n

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_m = n_m) = \frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

The product of powers of the p_i represents the probability of any *particular* sequence of results with n_i results in category i for each $1 \le i \le m$, while the ratio of factorials

$$\frac{n!}{n_1! n_2! \cdots n_m!} = \binom{n}{n_1}\binom{n - n_1}{n_2} \cdots \binom{n - n_1 - \cdots - n_{m-1}}{n_m}$$

called a *multinomial coefficient* is the number of different possible arrangements of symbols in a row of symbols made from n_1 symbols 1, n_2 symbols 2, ..., and n_m symbols m. A symbol i at place j in the row represents a result in category i on trial j. The derivation of this formula parallels the derivation of the binomial formula in Section 2.1, which is the special case $m = 2$. The multinomial distribution provides a natural example of a joint distribution of m variables $N_1, \ldots, N_m$ that are not independent, due to the constraint that $N_1 + \cdots + N_m = n$.

Example 7. Fours, fives, and sixes.

Suppose a fair die is rolled 10 times, and the numbers of rolls of four, five, and six are recorded.

Solution. From the multinomial distribution for $n = 10$ trials, $m = 4$ categories ("four", "five", "six", and "other") with probabilities $1/6, 1/6, 1/6$ and $3/6$, the required probability is

$$P(N_{\text{four}} = 1, N_{\text{five}} = 2, N_{\text{six}} = 3, N_{\text{other}} = 4) = \frac{10!}{1!2!3!4!}\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^3\left(\frac{3}{6}\right)^4$$

Symmetry

Symmetry arguments often simplify probability calculations. The basic idea is to recognize when probabilities of different events must be equal by symmetry.

Symmetry about 0. The distribution of X is *symmetric about* 0 if

$$P(X = -x) = P(X = x) \qquad \text{for all } x$$

A histogram displaying the distribution of X is then symmetric about 0 in the usual sense of reflection through the vertical axis. Equivalently, since $P(X = -x) = P(-X = x)$ for all x, $P(-X = x) = P(X = x)$ for all x. That is to say

$$-X \text{ has the same distribution as } X$$

Then for all a

$$P(X \geq a) = P(-X \leq -a) = P(X \leq -a)$$

Here the first equality holds because the two events $(X \geq a)$ and $(-X \leq -a)$ are identical (multiplication by -1: note the reversal of the inequality). Also the probability $P(-X \leq -a)$ equals $P(X \leq -a)$ because any statement about $-X$ has the same probability as the corresponding statement about X, by the equality in distribution of $-X$ and X (change of variable principle).

Example 8. Symmetry about 0 for sums of independent random variables.
Let $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are independent, and each X_i has a distribution that is symmetric about 0.

Problem. Show for every a

$$P(S_n \leq -a) = P(S_n \geq a)$$

Solution. In other words, the problem is to show that the distribution of S_n is symmetric about 0. Since, by assumption, $-X_i$ has the same distribution as X_i, and the X_i are independent, it follows that $(-X_1, \ldots, -X_n)$ has the same joint distribution as $(X_1, \ldots, X_n)$. This uses the fact that functions of independent random variables

are independent (applied to $f(X_i) = -X_i$). Adding the coordinates of the two sequences $(-X_1, \ldots, -X_n)$ and $(X_1, \ldots, X_n)$ shows that $-S_n = (-X_1) + \cdots + (-X_n)$ has the same distribution as S_n. That is to say, the distribution of S_n is symmetric about 0.

Discussion. Note the use of the following form of the change of variable principle for sequences of random variables: if $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ have the same joint distribution, then $g(X_1, \ldots, X_n)$ and $g(Y_1, \ldots, Y_n)$ have the same distribution for any function g of n variables. For instance, $X_1 + \cdots + X_n$ and $Y_1 + \cdots + Y_n$ have the same distribution. This fact was used in the example for $Y_i = -X_i$. Note also how the reasoning did not involve any explicit summation of probabilities in the joint distribution of $(X_1, \ldots, X_n)$, which would be necessary to find a formula for the distribution of S_n. This is the point of a symmetry argument: to show two probabilities are equal without calculating either of them.

Symmetry about *b*. The distribution of a random variable Y with a finite number of numerical values is *symmetric about b* if

$$P(Y = b + x) = P(Y = b - x) \qquad \text{for all } x$$

Equivalently, the distribution of $Y - b$ is symmetric about 0. Then for every c

$$P(Y \le b - c) = P(Y \ge b + c)$$

Symmetry for a sum of independent random variables. If Y_i has distribution symmetric about b_i, and the Y_i are independent, then $Y_1 + \cdots + Y_n$ has distribution symmetric about $b_1 + \cdots + b_n$. This follows from the result of the previous example applied to $X_i = Y_i - b_i$.

Example 9. Sum of 101 random digits.

Let S_{101} denote the sum of 101 independent random digits, each picked uniformly at random from $\{0, 1, \ldots, 9\}$.

Problem. Find $P(S_{101} \le 454)$.

Solution. Here $S_{101} = Y_1 + \cdots + Y_{101}$ for Y_i that are independent, and the distribution of each Y_i is symmetric about $4\frac{1}{2}$. So the distribution of S_{101} is symmetric about $101 \times (4\frac{1}{2}) = 454.5$. Therefore

$$P(S_{101} \le 454) = P(S_{101} \le 454.5 - .5) = P(S_{101} \ge 454.5 + .5) = P(S_{101} \ge 455)$$

But since S_{101} has integer values, $P(S_{101} \le 454) + P(S_{101} \ge 455) = 1$, which forces $P(S_{101} \le 454) = \frac{1}{2}$.

Discussion. For S_n the sum of n digits the argument shows that the distribution of S_n is symmetric about $(4\frac{1}{2})n$ for every n. For odd n, say $n = 2m+1$, this symmetry can be used just as above to identify a probability in the distribution of S_{2m+1} that is exactly $1/2$:

$$P(S_{2m+1} \le 9m + 4) = P(S_{2m+1} \ge 9m + 5) = \tfrac{1}{2}$$

For odd n the histogram of S_n has bars of equal height at the integers $(4\frac{1}{2})n \pm 1/2$, $(4\frac{1}{2})n \pm 3/2, \ldots$, so the distribution splits perfectly into two equal halves. For even n the histogram of S_n has a bar exactly on the point of symmetry $(4\frac{1}{2})n$, and equal bars at $(4\frac{1}{2})n \pm 1$, $(4\frac{1}{2})n \pm 2, \ldots$. Then the distribution of S_n does not split into equal halves to the right and left of $(4\frac{1}{2})n$, because there is a lump of probability right on the point of symmetry which cannot be split in two. It can be shown that for even n the central probability $P[S_n = (4\frac{1}{2})n]$ is actually the largest individual probabilty in the distribution of S_n. It will be seen in Section 3.3 that for large n the distribution of S_n follows a normal curve very closely. This is similar to what happens for large n to the binomial $(n, 1/2)$ distribution of $X_1 + \cdots + X_n$ for X_i picked at random from $\{0, 1\}$. It follows that as in the binomial case, for large even n the distribution of the sum of n digits has central term $P[S_n = (4\frac{1}{2})n]$ that converges to zero very slowly, like a constant over $\sqrt{n}$. For very large $n = 2m$ this term can be ignored, so

$$P(S_{2m} \le 9m) = P(S_{2m} \ge 9m) \approx \tfrac{1}{2}$$

The approximate probability $\frac{1}{2}$ is less than the true probability by

$$P(S_{2m} = 9m)/2 \sim c/\sqrt{m}$$

where the constant c can be shown using the normal approximation to be equal to $1/\sqrt{33\pi}$, and "$\sim$" means that the ratio of the two sides tends to 1 as $m \to \infty$. (See Exercise 3.3.31).

Exercises 3.1

1. Let X be the number of heads in three tosses of a fair coin.

a) Display the distribution of X in a table. b) Find the distribution of $|X - 1|$.

2. Let X and Y be the numbers obtained in two draws at random from a box containing four tickets 1, 2, 3, and 4. Display the joint distribution table for X and Y:

a) for sampling with replacement; b) for sampling without replacement.

Calculate $P(X \le Y)$ from the table in each case.

3. Suppose a fair die is rolled twice. Let S be the sum of the numbers on the two rolls.

a) What is the range of S? b) Find the distribution of S.

4. Let X_1 and X_2 be the numbers obtained on two rolls of a fair die. Let $Y_1 = \max(X_1, X_2)$, $Y_2 = \min(X_1, X_2)$. Display joint distribution tables for a) (X_1, X_2); b) (Y_1, Y_2).

5. Find the distribution of X_1X_2 for X_1 and X_2 as in Exercise 4.

6. A fair coin is tossed three times. Let X be the number of heads on the first two tosses, Y the number of heads on the last two tosses.

a) Make a table showing the joint distribution of X and Y.

b) Are X and Y independent? c) Find the distribution of $X + Y$.

7. Let A, B, and C be events that are independent, with probabilities a, b, and c. Let N be the random number of events that occur.

a) Express the event $(N = 2)$ in terms of A, B, and C. b) Find $P(N = 2)$.

8. A hand of five cards contains two aces and three kings. The five cards are shuffled and dealt one by one, until an ace appears.

a) Display in a table the distribution of the number of cards dealt.

b) Suppose that dealing is continued until the second ace appears. Again display the distribution of the number of cards dealt.

c) Explain why the probabilities in the second table are just those in the first in a different order. (*Hint*: Think about dealing off the bottom of the deck!)

9. A box contains 8 tickets. Two are marked 1, two marked 2, two marked 3, and two marked 4. Tickets are drawn at random from the box without replacement until a number appears that has appeared before. Let X be the number of draws that are made. Make a table to display the probability distribution of X.

10. Blocks of Bernoulli trials. In $n + m$ independent Bernoulli (p) trials, let S_n be the number of successes in the first n trials, T_m the number of successes in the last m trials.

a) What is the distribution of S_n? Why?

b) What is the distribution of T_m? Why?

c) What is the distribution of $S_n + T_m$? Why?

d) Are S_n and T_m independent? Why?

11. Binomial sums. Let U_n have binomial(n, p) distribution and let V_m have binomial(m, p) distribution. Suppose U_n and V_m are independent.

a) Find the distribution of $U_n + V_m$ without calculation by a simple argument that refers to the solution of Exercise 10.

b) Compare the result of part a) to a calculation of $P(U_n+V_m = k)$ for $0 \le k \le n+m$ from the joint distribution of U_n and V_m, and hence prove the identity

$$\sum_{j=0}^{n} \binom{n}{j}\binom{m}{k-j} = \binom{n+m}{k}$$

c) Derive the identity in part b) by a counting argument. [*Hint*: Classify the subsets of size k of $\{1, \ldots, n + m\}$ by how many elements of $\{1, \ldots, n\}$ they contain.]

d) Derive the identity in part b) in another way by finding the coefficient of $p^k q^{n+m-k}$ in $(p + q)^{n+m} = (p + q)^n(p + q)^m$ in two different ways.

e) Simplify the sum $\sum_{j=0}^{n} \binom{n}{j}^2$.

12. Grouping multinomial categories. Suppose that counts $(N_1, \ldots, N_m)$ are the numbers of results in m categories in n repeated trials. So $(N_1, \ldots, N_m)$ has multinomial distribution with parameters n and $p_1, \ldots, p_m$, as in the box above Example 7. Let $1 \le i < j \le m$. Answer the following questions with an explanation, but no calculation.

a) What is the distribution of N_i? b) What is the distribution of $N_i + N_j$?

c) What is the joint distribution of N_i, N_j, and $n - N_i - N_j$?

13. A box contains $2n$ balls of n different colors, with 2 of each color. Balls are picked at random from the box with replacement until two balls of the same color have appeared. Let X be the number of draws made.

a) Find a formula for $P(X > k)$, $k = 2, 3, \ldots$

b) Assuming n is large, use an exponential approximation to find a formula for k in terms of n such that $P(X > k)$ is approximately $1/2$. Evaluate k for n equal to one million.

14. In a World Series, teams A and B play until one team has won four games. Assume that each game played is won by team A with probability p, independently of all previous games.

a) For $g = 4$ through 7, find a formula in terms of p and $q = 1 - p$ for the probability that team A wins in g games.

b) What is the probability that team A wins the World Series, in terms of p and q?

c) Use your formula to evaluate this probability for $p = 2/3$.

d) Let X be a binomial $(7, p)$ random variable. Explain why $P(\text{A wins}) = P(X \ge 4)$ using an intuitive argument. Verify algebraically that this is true.

e) Let G represent the number of games played. What is the distribution of G? For what value of p is G independent of the winner of the series?

15. Let X and Y be independent, each uniformly distributed on $\{1, 2, \ldots, n\}$. Find:

a) $P(X = Y)$; b) $P(X < Y)$; c) $P(X > Y)$;

d) $P(\max(X, Y) = k)$ for $1 \le k \le n$;

e) $P(\min(X, Y) = k)$ for $1 \le k \le n$; f) $P(X + Y = k)$ for $2 \le k \le 2n$.

16. Discrete convolution formula. Let X and Y be independent random variables with non-negative integer values. Show that:

a) $P(X + Y = n) = \sum_{k=0}^{n} P(X = k)P(Y = n - k)$.

b) Find the probability that the sum of numbers on four dice is 8, by taking X to be the sum on two of the dice, Y the sum on the other two.

17. Let X be the number of heads in 20 fair coin tosses, Y a number picked uniformly at random from $\{0, 1, \ldots, 20\}$, independently of X. Let $Z = \max(X, Y)$.

a) Find a formula for $P(Z = k)$, $k = 0, \ldots, 20$.

b) Without calculating out $P(Z = k)$ exactly, sketch the histogram of Z, and explain its unusual shape.

18. Three dice are rolled.

a) What is the probability that the total number of spots showing is 11 or more? [*Hint*: No long calculations!]

b) Find a number m such that if five dice are rolled, the probability that the total number of spots showing is m or more is the same as this probability of 11 or more spots from three dice.

19. Sum of biased dice. Let S be the sum of numbers obtained by rolling two biased dice with possibly different biases described by probabilities $p_1, \ldots, p_6$, and $r_1, \ldots, r_6$, all assumed to be nonzero.

a) Find formulae for $P(S = k)$ for $k = 2$, 7, and 12.

b) Show that $P(S = 7) > P(S = 2)\frac{r_6}{r_1} + P(S = 12)\frac{r_1}{r_6}$.

c) Deduce that no matter how the two dice are biased, the numbers 2, 7, and 12 cannot be equally likely values for the sum. In particular, the sum cannot be uniformly distributed on the numbers from 2 to 12.

d) Do there exist positive integers a and b and independent non-constant random variables X and Y such that $X+Y$ has uniform distribution on the set of integers $\{a, a+1, \ldots, a+b\}$?

20. Pairwise independence. Let $X_1, \ldots, X_n$ be a sequence of random variables. Suppose that X_i and X_j are independent for every pair (i, j) with $1 \leq i < j \leq n$. Does this imply $X_1, \ldots, X_n$ are independent? Sketch a proof or counterexample.

21. Sequential independence. Let $X_1, \ldots, X_n$ be a sequence of random variables. Suppose that for every $1 \leq m \leq n-1$ the random sequence $(X_1, \ldots, X_m)$ is independent of the next random variable X_{m+1}. Does this imply $X_1, \ldots, X_n$ are independent? Sketch a proof or give a counterexample.

22. Suppose that random variables X and Y, each with a finite number of possible values, have joint probabilities of the form

$$P(X = x, Y = y) = f(x)g(y)$$

for some functions f and g, for all (x, y).

a) Find formulae for $P(X = x)$ and $P(Y = y)$ in terms of f and g.

b) Use your formulae to show that X and Y are independent.

23. Suppose X and Y are two random variables such that $X \geq Y$.

a) For a fixed number T, which would be greater, $P(X \leq T)$ or $P(Y \leq T)$?

b) What if T is a random variable?

24. Suppose a box contains tickets, each labeled by an integer. Let X, Y, and Z be the results of draws at random with replacement from the box: Show that, no matter what the distribution of numbers in the box,

a) $P(X + Y \text{ is even}) \geq 1/2$; b) $P(X + Y + Z \text{ is a multiple of } 3) \geq 1/4$.

3.2 Expectation

The *mean* or *expected value* of a random variable X is a number derived from the distribution of X the same way that the *mean* or *average* $\bar{x}$ of a list of numbers $(x_1, \ldots, x_n)$ is derived from the empirical distribution of the list:

$$\bar{x} = (x_1 + \cdots + x_n)/n = \sum_{\text{all } x} x P_n(x) \tag{1}$$

where $P_n(x)$ is the proportion of the n values x_k that are equal to x. These proportions $P_n(x)$, which sum to 1 over all x, define the empirical distribution of the list (see the end of Section 1.3). To illustrate, the average of the list $(1, 0, 8, 6, 6, 1, 6)$ of $n = 7$ numbers is

$$(1+0+8+6+6+1+6)/7 = 0 \times \frac{1}{7} + 1 \times \frac{2}{7} + 6 \times \frac{3}{7} + 8 \times \frac{1}{7} = 4$$

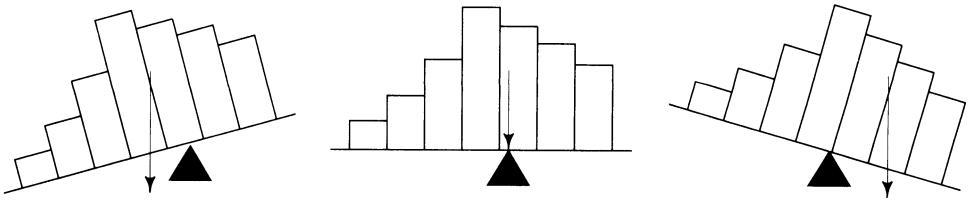
The second formula for $\bar{x}$ in (1) is a *weighted average* of values x with weights $P_n(x)$. This formula is obtained in general, just as in the example, by grouping terms with a common x-value. The weighted average formula for $\bar{x}$ suggests the following definition:

Mean of a Distribution

The *mean* μ of a probability distribution $P(x)$ over a finite set of numerical values x is the average of the values x weighted by their probabilities:

$$\mu = \sum_{\text{all } x} x\, P(x)$$

The center of gravity. If you think of a distribution of mass instead of probability, the mean is the *center of gravity*. Think of a histogram of the distribution as a shape cut from a rigid material of constant thickness and density. The mean value is then a balance point for the histogram. The shape balances when supported at the mean, tips over to the right when supported at a point to the left of the mean, and tips to the left when supported to the right of the mean. This is due to the principle of moments in mechanics.



Mean of the binomial distribution. It is shown later in this section that the general definition of the mean μ of a distribution is consistent with the formula $\mu = np$

for the binomial (n, p) distribution, used in Chapter 2. In n independent trials with probability p of success on each trial, you expect to get around $\mu = np$ successes. So it is natural to say that the expected number of successes in n trials is np. This suggests the following definition of the expected value $E(X)$ of a random variable X. For X the number of successes in n trials, this definition makes $E(X) = np$. See Example 7.

Definition of Expectation

The *expectation* (also called *expected value*, or *mean*) of a random variable X, is the mean of the distribution of X, denoted $E(X)$. That is

$$E(X) = \sum_{\text{all } x} x\, P(X = x)$$

the *average of all possible values of X, weighted by their probabilities.*

Example 1. Random sampling.

Suppose n tickets numbered $x_1, \ldots, x_n$ are put in a box and a ticket is drawn at random. Let X be the x-value on the ticket drawn. Then $E(X) = \bar{x}$, the ordinary average of the list of numbers in the box. This follows from the above definition, and the weighted average formula (1) for $\bar{x}$, because the distribution of X is the empirical distribution of x-values in the list:

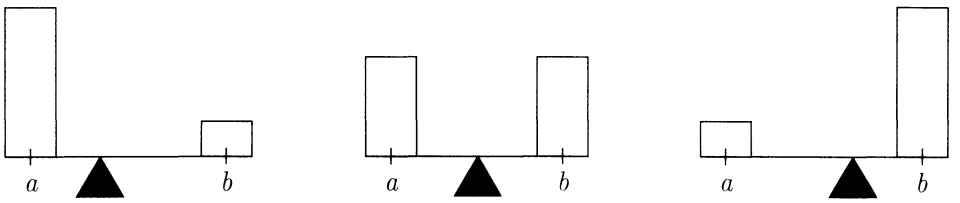
$$P(X = x) = P_n(x) = \#\{i : 1 \le i \le n \text{ and } x_i = x\}/n$$

Example 2. Two possible values.

If X takes two possible values, say a and b, with probabilities $P(a)$ and $P(b)$, then

$$E(X) = aP(a) + bP(b)$$

where $P(a) + P(b) = 1$. This weighted average of a and b is a number between a and b, proportion $P(b)$ of the way from a to b. The larger $P(a)$, the closer $E(X)$ is to a; and the larger $P(b)$, the closer $E(X)$ is to b.



Example 3. Indicators.

This is the special case of the previous example for $a = 0$ and $b = 1$. Suppose $X = I_A$ is the *indicator* of event A. Since I_A has value 1 if A occurs, 0 otherwise, the events $(I_A = 1)$ and A are identical by definition. So

$$E(I_A) = 0P(I_A = 0) + 1P(I_A = 1) = P(A)$$

Indicators may seem trivial at first. But they combine to produce more interesting random variables by sums and products. Examples follow later in this section.

Example 4. Rolling a die.

Suppose X is the number produced by rolling a fair die. The definition of $E(X)$ makes

$$\begin{aligned} E(X) &= 1P(X = 1) + 2P(X = 2) + \cdots + 6P(X = 6) \\ &= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5 \end{aligned}$$

Of course, you should not expect a single die roll to be 3.5. But if you roll the die a large number of times you should expect the average of the rolls to be close to 3.5. To see why, calculate the sum of the rolls by grouping terms of the same value:

$$\text{sum of the rolls} = 1 \times (\text{number of 1's}) + \cdots + 6 \times (\text{number of 6's})$$

Dividing by the total number of rolls now gives

$$\text{average of the rolls} = 1 \times (\text{proportion of 1's}) + \cdots + 6 \times (\text{proportion of 6's})$$

Assuming a large number of independent rolls, each of these proportions is likely to be very close to $1/6$, by the law of large numbers. The average of the rolls will then be close to $E(X) = 3.5$. If the die were biased, with probability p_i of rolling number i, the same reasoning shows the long-run average is likely to be very close to

$$E(X) = 1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6$$

The long-run interpretation of expectation. In general, the long-run argument in the last example leads to the conclusion in the next box. A more precise formulation of this idea, a law of averages for independent trials, is given in Section 3.3.

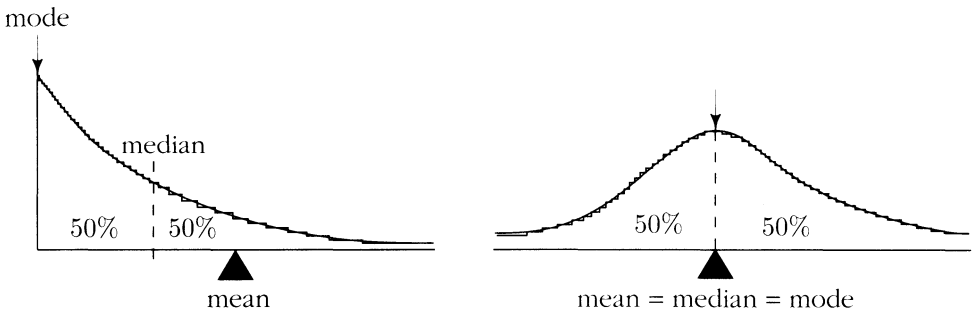
Expectation as a Long-Run Average

If probabilities for values of X are approximate long-run frequencies, then $E(X)$ is approximately the long-run average value of X.

Because expectation approximates a long-run average (and because Example 1 equates an expectation and an average), the properties of expectation described in this section parallel properties of the ordinary average of a list of numbers. A summary of these properties of averages and expectations is displayed on pages 180 – 181.

Comparison of the mean with other measures of location. The mean is one way to locate a central point in the distribution of X. But there are other ways, for example, the *mode* and the *median.* A mode is the most likely possible value of X (there may be more than one). And a median is a number m such that both $P(X \leq m)$ and $P(X \geq m)$ are at least $1/2$. There may be more than one median. For example, if X is the number on a fair die, every integer between 1 and 6 is a mode of X, and every number between 3 and 4 is a median of X. The mean, the mode, and the median may be quite different. But if the distribution is symmetric about some point m, and has a single mode, the three quantities all equal m. Of all measures of location, the mean is most important in theory. This is due to the close connection between means and long-run averages, and the fact shown later in this section that the mean of the sum of two random variables is the sum of the means. There is no such simple rule for modes or medians.

FIGURE 1. Mean, mode, and median.



Gambling Interpretation of Expectation: The Fair Price

Suppose you bet on an outcome of some kind. You pay a fixed amount $\$b$ to place the bet, and the return from the bet is the random amount $\$X$. For example, you might pay \$4 to buy a return of $\$X$ where X is the number produced by a fair die roll. Suppose you made a long series of such bets, with independent repetitions of whatever random mechanism generates X, for example successive rolls of the die, or successive spins of a roulette wheel. After n repetitions, you have paid out $\$nb$ to place the bets. The return from your bets is the sum $S_n = X_1 + \cdots + X_n$, where X_i is the return from the ith bet. The basic assumption is that the X_i are independent with the same distribution as X. By the law of large numbers, the long-run proportion of

trials that yield x is approximately $P(X = x)$. So over the n trials you should expect to see the return x about $nP(X = x)$ times. The total or *gross return* from a large number n of bets (not subtracting the price of the bets) should therefore be around

$$\$\sum_x x\, n\, P(X = x) = \$nE(X).$$

To summarize:

> *Over the long run, for a series of independent bets with returns like* $\$X$, *the average gross return per bet will probably be close to* $\$E(X)$.

If you pay the same price $\$b$ to bet each time, your long-run *net return per bet* from a large number of bets will probably be about $\$(E(X) - b)$. To illustrate, if you pay \$4 for the return of $\$X$ for X the number on a fair die roll, over the long run you should expect to lose about 50 cents a game. Such considerations lead to the following interpretation of $E(X)$ as a *fair price*:

> $\$E(X)$ *is the fair price to pay for a return of* $\$X$. *This price makes wins and losses tend to cancel out over the long run.*

Precise information about the degree of cancellation of wins and losses to be expected over the long run is provided by the normal approximation in the next section.

Indicator variables and fair odds. The idea of a fair price is a generalization of the fair odds rule presented in Section 1.1. Suppose you pay the price $\$b$ to get a return of \$1 if an event A occurs, and no return otherwise. The return from your bet is then $\$I_A$ where I_A is the indicator of A. The fair price for this return is $\$b$ where

$$b = E(I_A) = P(A).$$

This restates the fair odds rule (see Example 1.1.4).

The Addition Rule

Let $\$X$ and $\$Y$ be the returns from two bets on an outcome of some kind, for instance the returns from two stakes placed on different groups of numbers for a single spin of a roulette wheel. The combined return from the two bets is $\$(X+Y)$. It is quite intuitive that the fair price for this combination of two bets is

$$\$E(X+Y) = \$E(X) + \$E(Y),$$

the sum of the fair prices of the individual bets. This is the fundamental *addition rule* of expectation stated in the following box, and derived from the definition of expectation on page 177:

Addition Rule for Expectation

For any two random variables X and Y defined in the same setting,

$$E(X+Y) = E(X) + E(Y)$$

no matter whether X and Y are independent or not. Consequently, for a sequence of random variables $X_1, \ldots, X_n$, however dependent,

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$$

In calculations the definition of expectation

$$E(X) = \sum_{\text{all } x} x\, P(X = x)$$

is useful only if the formula for $P(X = x)$ allows an easy evaluation of the sum over all x of $xP(X = x)$. This happens only in the simplest examples. But even if the distribution of X is hard to compute, it is often possible to write X as a sum of simpler variables whose expectations are easily found. Then the expectation of X is found by the addition rule.

Example 5. Sum of dice.

Problem. Let T_n be the sum of numbers from n dice. Find $E(T_n)$.

Solution. Let $X_1, \ldots, X_n$ be the numbers obtained from the n die rolls. Then

$$\begin{aligned} T_n &= X_1 + \cdots + X_n, \qquad \text{so} \\ E(T_n) &= E(X_1) + \cdots + E(X_n) \qquad \text{by the addition rule} \\ &= 3.5 + \cdots + 3.5 \qquad (n \text{ terms}) \\ &= (3.5)n \end{aligned}$$

Discussion. Despite the fact that the distribution of T_n becomes more and more difficult to calculate exactly as n increases, the formula for $E(T_n)$ is simple. As a check, $E(T_2)$ can be found from its distribution:

$$\begin{aligned} E(T_2) = {} & 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + 5 \times \frac{4}{36} + 6 \times \frac{5}{36} + 7 \times \frac{6}{36} \\ & + 8 \times \frac{5}{36} + 9 \times \frac{4}{36} + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} \\ = {} & 7. \end{aligned}$$

The Method of Indicators

The idea of the method of indicators is that the random variable X that counts the number of events of some kind that occur can be represented as the sum of the indicators of these events. Then, by the addition rule for expectation, $E(X)$ is just the sum of the probabilities of the events. This is illustrated by the following two examples. First, it is worth restating the result of Example 3:

Expectation of an Indicator

The expectation of the indicator of an event is the probability of the event:

$$E(I_A) = P(A)$$

Example 6. Working components.

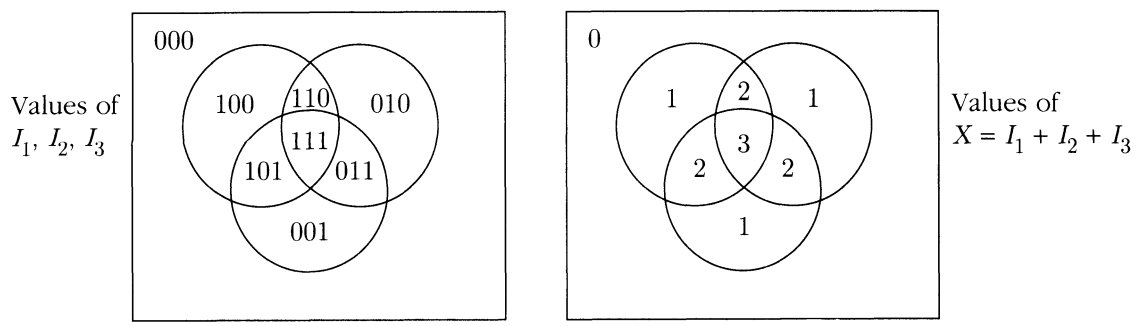
Suppose a system has n components, and that at a particular time the jth component is working with probability $p_j, j = 1, \ldots, n$. Let X be the number of components working at that time.

Problem. Find a formula for $E(X)$.

Solution. No matter which components work and which do not, the total number X that work can be found by adding 1 for each component that works and 0 for each component that does not. This is an expression for X in terms of indicators. Let I_j be the indicator random variable, which is 1 if the jth component is working, 0 otherwise. Then, as illustrated in Figure 2 for the case $n = 3$,

$$X = I_1 + I_2 + \cdots + I_n$$

FIGURE 2. Venn diagram for the number of working components. Here $n = 3$. The event that a particular component works is represented by the area inside a circle. These can overlap in any way.



Now take expectations of both sides. By the addition rule, and the fact that the expectation of I_j is p_j,

$$E(X) = p_1 + p_2 + \cdots + p_n$$

Discussion. You might think this problem could not be solved without further assumptions. True, the distribution of X cannot be found without assumptions about the dependence between the components. But due to the addition rule, $E(X)$ is the same, no matter what the dependence.

Example 7. Mean of the binomial distribution.

Suppose X is the number of successes in n independent trials with probability p of success in each trial, so X has binomial (n, p) distribution, as in Chapter 2.

Problem. Derive the formula $\mu = np$ for the mean of the binomial (n, p) distribution from the general definition of mean in this section.

Solution. As in the previous example, the total number of successes in the n trials can be written as a sum of indicators $X = I_1 + \cdots + I_n$ where I_j is the indicator of success on trial j, so $E(I_j) = p$ for each j, and the expected number of successes is

$$\begin{aligned} E(X) &= p + p + \cdots + p \qquad (n \text{ terms}) \\ &= np \end{aligned}$$

Discussion. This is not so obvious from the definition of $E(X)$:

$$E(X) = \sum_{\text{all } x} xP(X = x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

The calculation by the method of indicators implies that this expression must simplify to np. You can check this by algebra using the binomial theorem.

The general method. Examples 6 and 7 both illustrate the method of indicators. The general idea is that a random variable X with possible values $\{0, 1, \ldots, n\}$ can always be represented as counting the number of events that occur in some list of n events, say $A_1, \ldots, A_n$. Then X is called a *counting variable*. A suitable definition of the events A_j is usually clear from a verbal description of X. For instance,

- if X is the number of components that work among n components, let A_j be the event that the jth component works (Example 6).
- if X is the number of successes in n trials, let A_j be the event of success on trial j, for $1 \le j \le n$ (Example 7).
- if X is the number of aces in a 5-card poker hand, let A_j be the event that the jth card dealt is an ace, $1 \le j \le 5$ (Example 8).

The statement

$$X \text{ is the number of events } A_j \text{ that occur}$$

is expressed mathematically by the identity of random variables

$$X = I_1 + I_2 + \cdots + I_n \tag{2}$$

where I_j is the indicator of A_j. To illustrate, for X the number of aces in 5 cards, if the first and third cards are aces and the rest are not, this equation reads

$$2 = 1 + 0 + 1 + 0 + 0$$

while if the first three cards are aces and the last two are not, the equation reads

$$3 = 1 + 1 + 1 + 0 + 0$$

The point is that the number of aces can be found this way by adding zeros and ones, *no matter what the arrangement of the cards.* An equality like this, that holds by definition of the variables involved no matter what the outcome, is an *identity of random variables.* Take the expectation of both sides of (2), use the addition rule for expectation, and the fact that $E(I_j) = P(A_j)$ by definition of I_j as the indicator of A_j, to obtain the following generalization of the result of Examples 6 and 7:

Expected Number of Events that Occur

If X is the number of events that occur among some collection of events $A_1, \ldots, A_n$, then

$$E(X) = P(A_1) + P(A_2) + \cdots + P(A_n) \tag{3}$$

Usually it is easy to find $P(A_j)$, and add the results to find $E(X)$, as in Examples 6 and Example 7.

Example 8. The number of aces.

Let X be the number of aces in a 5-card poker hand. The probability that any particular card is an ace is 4/52 (Examples 1.4.7 and 3.1.6), so the expected number of aces among 5 cards dealt from a well-shuffled deck is

$$E(X) = 5 \times 4/52 = 5/13$$

Compare with the method of computing $E(X)$ directly from the definition of $E(X)$ in terms of the distribution of X which was found in Section 2.5:

$$P(X = x) = \sum_{x=0}^{4} xP(X = x) = \sum_{x=0}^{4} x \binom{4}{x}\binom{48}{5-x}\Big/\binom{52}{5}$$

You can check the second method gives the same answer. But the first method is quicker.

When to use the method of indicators. The examples show how the method of indicators can be used to find $E(X)$ for a counting variable X in either of the following circumstances:

- The probabilities $P(X = x)$ are known, but given by a formula that makes the expression $E(X) = \sum_x xP(X = x)$ hard to simplify.
- The nature of the dependence between the events A_j is either unknown, or known but so complicated that it is difficult to obtain a formula for $P(X = x)$.

The exact distribution of X depends in a fairly complicated way on the probabilities of various intersections of events being counted (Review Exercise 35). But, no matter what the dependence, the mean of the distribution is always given by the simple formula (3) for $E(X)$. There is usually more than one way to write a counting variable X as the sum of indicators of some collection of events. To find $E(X)$, all you need is one such collection of events whose probabilities you can calculate.

The tail sum formula for expectation of a counting variable. Every random variable with possible values $\{0, 1, \ldots, n\}$, however defined, is a counting variable representing number of events that occur in some list of n events $A_1, \ldots, A_n$. To see this, let A_j be the event $(X \geq j)$. If $X = x$ for $0 \leq x \leq n$, then A_j occurs for $1 \leq j \leq x$, and A_j does not occur for $x < j \leq n$. So if $X = x$ the number of events A_j that occur is precisely x. The resulting formula for $E(X)$ obtained by the method of indicators is displayed in the following box. Example 9 below gives an application.

Tail Sum Formula for Expectation

For X with possible values $\{0, 1, \ldots, n\}$,

$$E(X) = \sum_{j=1}^{n} P(X \geq j)$$

Alternative proof of the tail sum formula. Define $p_j = P(X = j)$. Then the expectation $E(X) = 1p_1 + 2p_2 + 3p_3 + \cdots + np_n$ is the following sum:

$$\begin{aligned} & p_1 \\ + & p_2 + p_2 \\ + & p_3 + p_3 + p_3 \\ & \cdots \quad \cdots \quad \cdots \quad \cdots \\ + & p_n + p_n + p_n + \cdots + p_n \end{aligned}$$

By the addition rule of probabilities, and the assumption that the only possible values of X are $\{0, 1, \ldots, n\}$, the sum of the first column of p's is $P(X \geq 1)$, the sum of the second column is $P(X \geq 2)$, and so on. The sum of the jth column is $P(X \geq j)$, $1 \leq j \leq n$. The whole sum is the sum of the column sums. □

Example 9. Expectation of a minimum.

Suppose that four dice are rolled.

Problem 1. Let M be the minimum of four numbers rolled. Find $E(M)$.

Solution. For any $1 \leq j \leq 6$, the event $(M \geq j)$ means that each X_i is at least j, where X_i is the number on the ith die. Thus

$$P(M \geq j) = P(X_1 \geq j, X_2 \geq j, X_3 \geq j, X_4 \geq j) = \left(\frac{6 - j + 1}{6}\right)^4$$

by independence of the X's, and fact that there are $6 - j + 1$ possible values for each X between j and 6. The tail sum formula gives

$$\begin{aligned} E(M) &= P(M \geq 1) + P(M \geq 2) + \cdots + P(M \geq 6) \\ &= \left(\frac{6}{6}\right)^4 + \left(\frac{5}{6}\right)^4 + \left(\frac{4}{6}\right)^4 + \left(\frac{3}{6}\right)^4 + \left(\frac{2}{6}\right)^4 + \left(\frac{1}{6}\right)^4 \approx 1.755 \end{aligned}$$

Discussion. The point of using the tail sum formula in this example is that the tail probabilities $P(M \geq j)$ are simpler than the individual probabilities

$$P(M = m) = P(M \geq m) - P(M \geq m + 1)$$

If you substitute this in the definition $E(M) = \sum_m mP(M = m)$, and simplify, you will find the coefficient of $P(M \geq j)$ is 1 for each j from 1 to n. That is the substance of the tail sum formula.

Problem 2. Let S be the sum of the largest three numbers among four dice. Find $E(S)$.

Solution. Notice that $S = T - M$, where T is the sum of all four numbers, and M is the minimum number. From Example 5, $E(T) = 4 \times (3.5) = 14$, and the value of $E(M)$ was just found. Since by the addition rule for expectation,

$$E(T) = E(T - M) + E(M) = E(S) + E(M)$$

$$E(S) = E(T) - E(M) = 14 - 1.755 = 12.245$$

Remark. It is much harder to find $E(S)$ via the distribution of S.

When is the sum of indicators an indicator? A sum of 0's and 1's is 0 or 1 if and only if there is at most a single 1 among all the terms. For events A_j with indicators I_j, this means that $\sum_j I_j$ is an indicator variable if and only if at most one of the events A_j can occur, that is, if and only if the events A_j are *mutually exclusive*. Then $\sum_j I_j$ is the indicator of the event $\bigcup_j A_j$ that at least one of the events A_j occurs. So in this case the result of the method of indicators is just the addition rule for probabilities:

$$P(\bigcup_j A_j) = \sum_j P(A_j) \text{ if the } A_j \text{ are mutually exclusive.}$$

Boole's inequality. In general, for possibly overlapping events A_j, the above equality is replaced by *Boole's inequality* of Exercise 1.3.13:

$$P(\bigcup_j A_j) \leq \sum_j P(A_j)$$

If X is the number of events A_j that occur, the left side is $P(X \geq 1)$, and the right side is $E(X)$. So Boole's inequality can be restated as follows: for any counting random variable X,

$$P(X \geq 1) \leq E(X)$$

This follows from the addition rule of probabilities and the definition of $E(X)$:

$$\begin{aligned} P(X \geq 1) &= p_1 + p_2 + p_3 + \cdots + p_n \\ &\leq p_1 + 2p_2 + 3p_3 + \cdots + np_n = E(X) \end{aligned}$$

To illustrate, Example 8 showed the expected number of aces among 5 cards is $5/13$. So the probability of at least one ace among 5 cards is at most $5/13 \approx 0.385$. The exact probability of at least one ace among 5 cards is $1 - \binom{48}{5}/\binom{52}{5} \approx 0.341$. In this case the upper bound of Boole's inequality is quite close to the exact probability of the union of events, because the probability of two or more aces in 5 cards is rather small (about 0.042). In other words, the events $A_1, \ldots, A_5$ do not overlap very much.

A generalization of Boole's inequality, called Markov's inequality, is illustrated by the following example:

Example 10. Bounding a tail probability.

Problem. For a non-negative random variable X with mean $E(X) = 3$, what is the largest that $P(X \geq 100)$ could possibly be?

Solution. The constraint that X is non-negative, i.e., $X \geq 0$, means that $P(X \geq 0) = 1$. In other words, all the probability in the distribution of X is in the interval $[0, \infty)$. Think of balancing a distribution of mass at 3, with all the mass in $[0, \infty)$. How can you get as much mass as possible in the interval $[100, \infty)$? Intuitively, the best you can do is to put some of the mass at 100 and the rest at 0 (as far to the left as allowed by the non-negativity constraint). This distribution balances at 3 if the proportion at 100 is 3/100. This shows $P(X \geq 100)$ can be as large as 3/100, and suggests it cannot be larger. Here is a proof. In the sum

$$\sum_{\text{all } x} xP(X = x) = 3$$

the terms with $x \geq 100$ contribute

$$\sum_{x \geq 100} xP(X = x) \geq \sum_{x \geq 100} 100P(X = x) = 100P(X \geq 100)$$

while all the terms are non-negative by the assumption that $X \geq 0$. This then gives $3 \geq 100P(X \geq 100)$, or $P(X \geq 100) \leq 3/100$.

Discussion. With arbitrary $E(X)$ and a instead of 3 and 100, this proves the following inequality. The point is that if $X \geq 0$, meaning all the possible values of X are non-negative, or $P(X \geq 0) = 1$, then knowing $E(X)$ puts a bound on how large the tail probability $P(X \geq a)$ can be.

Markov's Inequality

If $X \geq 0$, then $P(X \geq a) \leq \dfrac{E(X)}{a}$ for every $a > 0$.

Expectation of a Function of a Random Variable

Recall from Section 3.1 that if X is a random variable with a finite set of possible values, and $g(x)$ is a function defined on this set of possible values, then $g(X)$ is also a random variable. Examples of typical functions of a random variable X, whose expectations may be of interest, are X, X^2, X^k for some other power k, $\log(X)$ (assuming $X > 0$), e^X, or z^X for some other number z. The notation $g(X)$ is used for a generic function of X.

Expectation of a Function of X

Typically, $E[g(X)] \neq g[E(X)]$. Rather

$$E[g(X)] = \sum_{\text{all } x} g(x)P(X = x) \tag{4}$$

This formula is valid for any numerical function g defined on the set of possible values of X. In particular, for $g(x) = x^k$ with $k = 1, 2, \ldots$ the number

$$E(X^k) = \sum_{\text{all } x} x^k P(X = x)$$

derived from the distribution of X is called the *kth moment of* X.

The point of formula (4) is that it expresses $E[g(X)]$ directly in terms of the distribution of X, without consideration of the set of possible values of $g(X)$ or the distribution of $g(X)$ over these values. This is an important shortcut in many calculations.

Proof of the formula for $E[g(x)]$. Look at the sum $\sum_{\text{all } x} g(x)P(X = x)$, which is claimed to equal $E[g(X)]$. Group the terms according to the value y of $g(x)$. The terms from x with $g(x) = y$ have sum

$$\sum_{x:g(x)=y} g(x)P(X = x) = \sum_{x:g(x)=y} yP(X = x) = yP(g(X) = y)$$

Now summing over all y gives $E[g(X)]$. □

Constant factors. If X is a random variable, then so is cX for any constant c. This is $g(X)$ for $g(x) = cx$. Apply the formula for $E[g(X)]$ and factor the c out of the sum to see that $E(cX) = cE(X)$. So constants can be pulled outside the expectation operator.

Constant random variables. It is sometimes useful to think of a constant c as a random variable with just one possible value c. Of course, the expected value of a constant random variable is its constant value.

Linear functions. The expectation of a linear function of X is determined by the mean or first moment of X:

$$E(aX + b) = E(aX) + E(b) = aE(X) + b$$

This is immediate from the addition rule and the last two paragraphs. Linear functions $g(x) = ax + b$ are exceptional in that $E[g(X)] = g(E(X))$, a rule that is false for a general function g.

Moments. The *first moment* of X is just the mean or expectation of X. The *second moment* of X is $E(X^2)$, sometimes called the *mean square* of X. The term *moment* is borrowed from mechanics where similar averages with respect to a distribution of mass rather than probability have physical interpretations (principle of moments, moment of inertia). The moments of X are features of the distribution of X. Two random variables with the same distribution have the same moments. The first two moments of a distribution are by far the most important. The first moment gives a central value in the distribution. It will be seen in the next section that a quantity called *variance* derived from the first two moments gives an indication of how spread out the distribution is. Third moments are used to describe the degree of asymmetry of a distribution. Higher moments of X are hard to interpret intuitively. But they play an important part in theoretical calculations beyond the scope of this book.

It will be seen in the next section that

$$E(X^2) \neq [E(X)]^2$$

except in the trivial case when X is a constant random variable.

Example 11. Uniform distribution on three values.

If X is uniformly distributed on $\{-1, 0, 1\}$, then X has mean

$$E(X) = -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0$$

so $[E(X)]^2 = 0$. But, by the formula for $E[g(X)]$ with $g(X) = X^2$, the second moment of X is

$$E(X^2) = (-1)^2 \times \frac{1}{3} + 0^2 \times \frac{1}{3} + 1^2 \times \frac{1}{3} = \frac{2}{3} \neq [E(X)]^2 = 0$$

Quadratic functions. The first two moments of X determine the expectation of any quadratic function of X. For instance, the quantity $E[(X - b)^2]$ for a constant b, which arises in a prediction problem considered below, is found by expanding $(X - b)^2 = X^2 - 2bX + b^2$ and using the rules of expectation to obtain

$$E[(X - b)^2] = E[X^2 - 2bX + b^2] = E(X^2) - 2bE(X) + b^2$$

Functions of two or more random variables. The proof of the formula for $E[g(X)]$ shows that this formula is valid for any numerical function g of a random variable X with a finite number of possible values, even if these values are not numerical. In particular, substituting a random pair (X, Y) instead of X gives a formula for the expectation of $g(X, Y)$ for a generic numerical function g of two variables:

$$E[g(X, Y)] = \sum_{\text{all } (x,y)} g(x, y) P(X = x, Y = y)$$

Proof of the addition rule. Think of X, Y, and $X+Y$ as three different functions of X and Y, two random variables with a joint distribution specified by probabilities $P(x,y) = P(X = x, Y = y)$. By three applications of the formula for $E[g(X,Y)]$,

$$E(X) = \sum_{\text{all } (x,y)} xP(x,y)$$

$$E(Y) = \sum_{\text{all } (x,y)} yP(x,y)$$

$$E(X+Y) = \sum_{\text{all } (x,y)} (x+y)P(x,y)$$

Add the expressions for $E(X)$ and $E(Y)$ and simplify to get the expression for $E(X+Y)$. Conclusion: The addition rule $E(X) + E(Y) = E(X+Y)$.

Expectation of a product. As in the proof of the addition rule, view XY, the product of X and Y, as a function of (X,Y) to obtain

$$E(XY) = \sum_x \sum_y xyP(X = x, Y = y)$$

where the double sum is a sum over all pairs (x,y) of possible values for (X,Y). This formula holds regardless of whether or not X and Y are independent. If X and Y are independent, the formula can be simplified as follows:

$$E(XY) = \sum_x \sum_y xyP(X = x)P(Y = y)$$

$$= \left[\sum_x xP(X = x)\right]\left[\sum_y yP(Y = y)\right] = [E(X)][E(Y)]$$

This yields the following:

Multiplication Rule for Expectation

If X and Y are **independent** then

$$E(XY) = [E(X)][E(Y)]$$

This multiplication rule will be used in the next section. Note well the assumption of independence. In contrast to the addition rule, the multiplication rule does not hold in general for dependent random variables. For example, if $X = Y$, the left side becomes $E(X^2)$ and the right side becomes $[E(X)]^2$. These two quantities are typically not equal (Example 11).

Expectation and Prediction

Suppose you want to predict the value of a random variable X. What is the best predictor of X? To define "best" you must decide on a criterion and a class of predictors. The simplest prediction problem is to predict the value of X by a constant, say b. Think in terms of losing some amount $L(x,b)$ if you predict b and the value of X is actually x. The function $L(x,b)$ is called a *loss function* in decision theory. It seems reasonable to try to pick b so as to minimize the *expected loss*, or *risk* $r(b) = E[L(X,b)]$

Example 12. **Right or wrong.**

Suppose that $L(x,b) = 0$ if $x = b$, and 1 otherwise. So you are penalized nothing if you get the value of X right, and penalized by one unit if you get the value of X wrong.

Problem. What is the best predictor?

Solution. $E[L(X,b)] = 0P(X = b) + 1P(X \neq b) = 1 - P(X = b)$.
So choosing b to minimize expected loss for this loss function is the same as choosing b to maximize $P(X = b)$. That is to say, b should be a mode of the distribution of X. Many probability distributions have a unique mode. But every possible value of a uniformly distributed random variable is a mode.

Example 13. **Absolute error.**

Suppose $L(x,b) = |x - b|$. So the penalty is the absolute value of the difference between the actual value and the predicted value. Now there is a bigger penalty for bigger mistakes. The expected loss is

$$r(b) = E(|X - b|) = \sum_x |x - b| P(X = x)$$

by the formula for $E[g(X)]$ applied to $g(x) = |x - b|$ for fixed b.

Problem. Find b that minimizes $r(b)$.

Solution. This time the solution is the median. To see why, look for a fixed x at the derivative

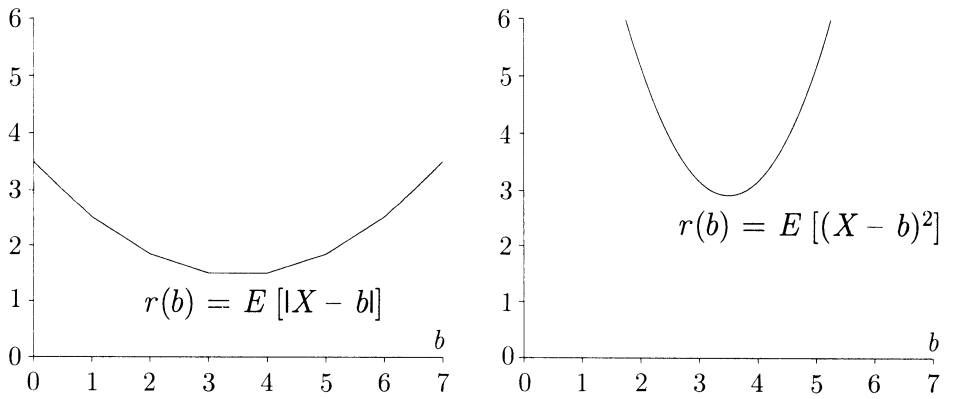
$$\frac{d}{db}|x - b| = \begin{cases} -1 & \text{if } b < x \\ 1 & \text{if } b > x \end{cases}$$

The sum defining $r(b)$ is over all possible values of X, say $x_1 < x_2 < \cdots < x_n$. So provided that $b \neq x_k$ for any k, the function $r(b)$ has the derivative

$$\begin{aligned}\frac{dr(b)}{db} &= \sum_{x<b} 1P(X=x) + \sum_{x>b}(-1)P(X=x) \\ &= P(X<b) - P(X>b) \\ &= 2P(X \le x_k) - 1 \qquad \text{if} \quad x_k < b < x_{k+1}\end{aligned}$$

So the function $r(b)$ is piecewise linear for b between x_k and x_{k+1}, decreasing if $P(X \le x_k) < 1/2$, increasing if $P(X \le x_k) > 1/2$, and flat if $P(X \le x_k) = 1/2$. So a b is minimizing if and only if $P(X \le b) \ge 1/2$ and $P(X \ge b) \ge 1/2$. Such a value b is a *median* of the distribution of X. A median always exists, but it may not be unique.

FIGURE 3. Risk functions for a die roll X with uniform distribution on $\{1, \ldots, 6\}$. **Left:** Graph of the risk function $r(b) = E(|X - b|)$ for absolute error. (Refer to Example 13.) In this example, every number in the interval $[3, 4]$ is a median for X. Numbers in this interval are equally good as predictors of X according to the criterion of minimizing the expected absolute error, and better than any other number. **Right:** The risk function $r(b) = E[(X-b)^2]$ for quadratic loss function. (Refer to Example 14.) Now $E(X) = 3.5$ is the unique best predictor.



Example 14. Squared error.

Suppose now the penalty is *squared error*, using the *quadratic loss function* $L(x, b) = (x - b)^2$.

Problem. Find b that is the best constant predictor of X for this quadratic loss function.

Solution. This time the answer is just the mean. Now

$$\begin{aligned} r(b) &= E[(X-b)^2] = E(X^2) - 2bE(X) + b^2 \\ \frac{dr(b)}{db} &= -2E(X) + 2b \end{aligned}$$

so $b = E(X)$ gives the *unique* best predictor of X for the quadratic loss function.

Properties of Averages

Definition. The average of a list of numbers $x_1, \ldots, x_n$ is

$$\bar{x} = (x_1 + \cdots + x_n)/n = \sum_{\text{all } x} x P_n(x)$$

where $P_n(x)$ is the proportion of the n values x_k that are equal to x (empirical distribution of the list).

Constants. If $x_k = c$ for every k, then

$$\bar{x} = c$$

Indicators. If every number x_k in a list is either a zero or a one, then

$$\bar{x} = \text{proportion of ones in the list}$$

Functions. If $y_k = g(x_k)$ for each k, typically $\bar{y} \neq g(\bar{x})$. But

$$\bar{y} = \sum_{\text{all } x} g(x) P_n(x)$$

Constant factors. If $y_k = cx_k$ for every k, where c is constant, then

$$\bar{y} = c\bar{x}$$

Addition. If $s_k = x_k + y_k$ for each k, then

$$\bar{s} = \bar{x} + \bar{y}$$

Multiplication. If $z_k = x_k y_k$ for each k, typically $\bar{z} \neq \bar{x}\bar{y}$.

Properties of Expectation

Definition. The expectation of a random variable X is

$$E(X) = \sum_{\text{all } x} xP(X = x)$$

(average of values of X weighted by their probabilities).

Constants. The expectation of a constant random variable is its constant value

$$E(c) = c$$

Indicators. If I_A is the indicator of an event A, so $I_A = 1$ if A occurs, 0 otherwise, then

$$E(I_A) = P(A)$$

Functions. Typically, $E[g(X)] \neq g[E(X)]$, but

$$E[g(X)] = \sum_{\text{all } x} g(x)P(X = x)$$

Constant factors. For a constant c,

$$E(cX) = cE(X)$$

Addition. The expectation of a sum of random variables is the sum of the expectations:

$$E(X + Y) = E(X) + E(Y) \qquad \text{even if } X \text{ and } Y \text{ are dependent.}$$

Multiplication. Typically, $E(XY) \neq E(X)E(Y)$. But

$$E(XY) = E(X)E(Y) \qquad \text{if } X \text{ and } Y \text{ are independent.}$$

Exercises 3.2

1. Suppose that 10% of the numbers in a list are 15, 20% of the numbers are 25, and the remaining numbers are 50. What is the average of the numbers in the list?

2. One list of 100 numbers contains 20% ones and 80% twos. A second list of 100 numbers contains 50% threes and 50% fives. A third list is obtained by taking each number in the first list and adding the corresponding number in the second list.

a) What is the average of the third list? Or is this not determined by the information given?

Repeat a) with adding replaced by b) subtracting c) multiplying by d) dividing by.

3. What is the expected number of sixes appearing on three die rolls? What is the expected number of odd numbers?

4. Suppose all the numbers in a list of 100 numbers are non-negative, and the average of the list is 2. Prove that at most 25 of the numbers in the list are greater than 8.

5. In a game of Chuck-a-Luck, a player can bet \$1 on any one of the numbers $1, 2, 3, 4, 5$, and 6. Three dice are rolled. If the player's number appears k times, where $k \geq 1$, the player gets $\$k$ back, plus the original stake of \$1. Otherwise, the player loses the \$1 stake. Some people find this game very appealing. They argue that they have a $1/6$ chance of getting their number on each die, so at least a $1/6+1/6+1/6 = 50\%$ chance of doubling their money. That's enough to break even, they figure, so the possible extra payoff in case their number comes up more than once puts the game in their favor.

a) What do you think of this reasoning?

b) Over the long run, how many cents per game should a player expect to win or lose playing Chuck-a-Luck?

6. Let X be the number of spades in 7 cards dealt from a well-shuffled deck of 52 cards containing 13 spades. Find $E(X)$.

7. In a circuit containing n switches, the ith switch is closed with probability p_i, $i = 1, \ldots, n$. Let X be the total number of switches that are closed. What is $E(X)$? Or is it impossible to say without further assumptions?

8. Suppose $E(X^2) = 3, \quad E(Y^2) = 4, \quad E(XY) = 2$. Find $E\left[(X+Y)^2\right]$.

9. Let X and Y be two independent indicator random variables, with $P(X = 1) = p$ and $P(Y = 1) = r$. Find $E[(X - Y)^2]$ in terms of p and r.

10. Let A and B be independent events, with indicator random variables I_A and I_B.

a) Describe the distribution of $(I_A + I_B)^2$ in terms of $P(A)$ and $P(B)$.

b) What is $E(I_A + I_B)^2$?

11. There are 100 prize tickets among 1000 tickets in a lottery. What is the expected number of prize tickets you will get if you buy 3 tickets? What is a simple upper bound for the probability that you will win at least one prize? Compare with the actual probability. Why is the bound so close?

12. Show that if a and b are constants with $P(a \leq X \leq b) = 1$, then $a \leq E(X) \leq b$.

13. Suppose a fair die is rolled ten times. Find numerical values for the expectations of each of the following random variables:

a) the sum of the numbers in the ten rolls;

b) the sum of the largest two numbers in the first three rolls;

c) the maximum number in the first five rolls;

d) the number of multiples of three in the ten rolls;

e) the number of faces which fail to appear in the ten rolls;

f) the number of different faces that appear in the ten rolls;

14. A building has 10 floors above the basement. If 12 people get into an elevator at the basement, and each chooses a floor at random to get out, independently of the others, at how many floors do you expect the elevator to make a stop to let out one or more of these 12 people?

15. Predicting demand. Suppose that a store buys b items in anticipation of a random demand Y, where the possible values of Y are non-negative integers y representing the number of items in demand. Suppose that each item sold brings a profit of $\$\pi$, and each item stocked but unsold brings a loss of $\$\lambda$. The problem is to choose b to maximize expected profit.

a) Show that this problem is the same as the problem of finding the predictor b of Y which minimizes over all integers the expected loss, with loss function

$$L(y, b) = \begin{cases} -\pi y + \lambda(b - y) & \text{if} \quad y \leq b \\ -\pi b & \text{if} \quad y > b \end{cases}$$

b) Let $r(b) = E[L(Y, b)]$. Use calculus to show that $r(b)$ is minimized over all the real numbers b, and hence over all the integers b, at the least integer y such that $P(Y \leq y) \geq \pi/(\lambda + \pi)$. *Note.* If $\pi = \lambda$, this is the median. If $\pi/(\lambda + \pi) = k\%$, this y is called the kth percentile of the distribution of Y.

16. Aces. A standard deck of 52 cards is shuffled and dealt. Let X_1 be the number of cards appearing before the first ace, X_2 the number of cards between the first and second ace (not counting either ace), X_3 the number between the second and third ace, X_4 the number between the third and fourth ace, and X_5 the number after the last ace. It can be shown that each of these random variables X_i has the same distribution, $i = 1, 2, \ldots, 5$, and you can assume this to be true.

a) Write down a formula for $P(X_i = k)$, $0 \leq k \leq 48$.

b) Show that $E(X_i) = 9.6$. [*Hint*: Do not use your answer to a).]

c) Are $X_1, \ldots, X_5$ pairwise independent? Prove your answer.

17. A box contains 3 red balls, 4 blue balls, and 6 green balls. Balls are drawn one-by-one without replacement until all the red balls are drawn. Let D be the number of draws made. Calculate: a) $P(D \leq 9)$; b) $P(D = 9)$; c) $E(D)$.

18. Suppose that X is a random variable with just two possible values a and b. For $x = a$ and b find a formula for $p(x) = P(X = x)$ in terms of a, b and $\mu = E(X)$.

19. A collection of tickets comes in four colors: red, blue, white, and green. There are twice as many reds as blues, equal numbers of blues and whites, and three times as many greens as whites. I choose 5 tickets at random with replacement. Let X be the number of different colors that appear.

a) Find a numerical expression for $P(X \geq 4)$.

b) Find a numerical expression for $E(X)$.

20. Show that the distribution of a random variable X with possible values 0, 1, and 2 is determined by $\mu_1 = E(X)$ and $\mu_2 = E(X^2)$, by finding a formula for $P(X = x)$ in terms of μ_1 and μ_2, $\quad x = 0, 1, 2$.

21. Indicators and the inclusion–exclusion formula. Let I_A be the indicator of A. Show the following:

a) the indicator of A^c, the complement of A, is $I_{A^c} = 1 - I_A$;

b) the indicator of the intersection AB of A and B is the product of I_A and I_B: $I_{AB} = I_A I_B$;

c) For any collection of events $A_1, \ldots, A_n$, the indicator of their union is

$$I_{A_1 \cup A_2 \cup \cdots \cup A_n} = 1 - (1 - I_{A_1})(1 - I_{A_2}) \cdots (1 - I_{A_n})$$

d) Expand the product in the last formula and use the rules of expectation to derive the inclusion–exclusion formula of Exercise 1.3.12.

22. Success runs in independent trials. Consider a sequence of $n \geq 4$ independent trials, each resulting in success (S) with probability p, and failure (F) with probability $1 - p$. Say a *run of three successes* occurs at the beginning of the sequence if the first four trials result in SSSF; a run of three successes occurs at the end of the sequence if the last four trials result in FSSS; and a run of three successes elsewhere in the sequence is the pattern FSSSF. Let $R_{3,n}$ denote the number of runs of three successes in the n trials.

a) Find $E(R_{3,n})$.

b) Define $R_{m,n}$, the number of success runs of length m in n trials, similarly for $1 \leq m \leq n$. Find $E(R_{m,n})$.

c) Let R_n be the total number of non-overlapping success runs in n trials, counting runs of any length between 1 and n. Find $E(R_n)$ by using the result of b).

d) Find $E(R_n)$ another way by considering for each $1 \leq j \leq n$ the number of runs that start on the jth trial. Check that the two methods give the same answer.

3.3 Standard Deviation and Normal Approximation

If you try to predict the value of a random variable X by its mean $E(X) = \mu$, you will be off by the random amount $X - \mu$. It is often important to have an idea of how large this deviation is likely to be. Because

$$E(X - \mu) = E(X) - \mu = 0$$

it is necessary to consider either the absolute value or the square of $X - \mu$ to get an idea of the size of the deviation without regard to sign. Because the algebra is easier with squares than with absolute values, it is natural to first consider $E[(X - \mu)^2]$, then take a square root to get back to the same scale of units as X.

Definition of Variance and Standard Deviation

The *variance* of X, denoted $Var(X)$, is the mean squared deviation of X from its expected value $\mu = E(X)$:

$$Var(X) = E[(X - \mu)^2]$$

The *standard deviation* of X, denoted $SD(X)$, is the square root of the variance of X:

$$SD(X) = \sqrt{Var(X)}$$

Intuitively, $SD(X)$ should be understood as a measure of how spread out the distribution of X is around its mean μ. Because $Var(X)$ is a central value in the distribution of $(X - \mu)^2$, its square root $SD(X)$ gives a rough idea of the typical size of the absolute deviation $|X - \mu|$. Variance always appears as an intermediate step in the calculation of standard deviation. Variance is harder to interpret than SD, but has simpler algebraic properties. Notice that $E(X)$, $Var(X)$, and $SD(X)$ are all determined by the distribution of X. That is to say, if two random variables have the same distribution, then they have the same mean, variance, and SD. So we may speak of the mean, variance, and SD of a distribution rather than a random variable.

Parameters of a normal curve. If a histogram displaying the distribution of X follows an approximately normal curve, the curve will be centered near the mean $E(X)$, and $SD(X)$ will be approximately the distance between the center of the curve and its shoulders, where the curve switches from being concave to convex. See Figure 1 of Section 2.2. This observation is justified at the end of Section 4.1. For histograms which are approximately normal in shape, about 68% of the probability will lie in the interval within one standard deviation of the mean.

Meaning of SD when the distribution is not roughly normal. If the distribution of X is not roughly normal, there is no simple way to visualize $SD(X)$ in terms of the histogram of X. But no matter what the distribution of X, you should expect X to be around $E(X)$, plus or minus a few times $SD(X)$. This is made more precise later in this section by Chebychev's inequality. Like the mean $E(X)$, the standard deviation $SD(X)$ can be interpreted in terms of a sum $S_n = X_1 + \cdots + X_n$ of a large number n of random variables X_i with the same distribution as X. What happens is that for large n the distribution of S_n follows an approximately normal curve with parameters determined by $E(X)$, $SD(X)$, and n. This is made precise by the *central limit theorem* stated later in this section.

It is often simpler to calculate an SD using the following formula for variance rather than the definition.

Computational Formula for Variance

$$Var(X) = E(X^2) - [E(X)]^2 = \sum_{\text{all } x} x^2 P(X = x) - \left[\sum_{\text{all } x} xP(X = x)\right]^2$$

In words: Variance is the mean of the square minus the square of the mean.

Remark. The order of the two operations, squaring and taking expectation, is extremely important. Since from its original definition $Var(X)$ is non-negative, and zero if and only if $P(X = \mu) = 1$, the computational formula shows that

$$E(X^2) \geq [E(X)]^2$$

with equality if and only if X is a constant random variable.

Proof.

$$\begin{aligned} E[(X-\mu)^2] &= E[X^2 - 2\mu X + \mu^2] \\ &= E(X^2) - 2\mu^2 + \mu^2 \qquad \text{by rules of } E \text{ using } E(X) = \mu \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2 \qquad \text{because } \mu = E(X) \end{aligned}$$

The second expression in the box comes from the formula for the expectation of a function of X, applied to $f(x) = x^2$, and the definition of $E(X)$. □

Example 1. Random sampling.

Suppose n tickets numbered $x_1, \ldots, x_n$ are put in a box and a ticket is drawn at random. Let X be the x-value on the ticket drawn. Then $E(X) = \bar{x}$, the average of the list of numbers in the box, as shown in Example 3.2.1. The corresponding formula for the standard deviation is $SD(X) = \sqrt{Var(X)}$ where

$$Var(X) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$$

The first formula comes from writing $X = x_I$ where I has uniform distribution on $\{1, 2, \ldots, n\}$, so $E[(X-\mu)^2] = E[(x_I - \bar{x})^2]$ is the expectation of a function of I. The second formula follows similarly from the computational formula for variance. The numbers $Var(X)$ and $SD(X)$ determined this way by a list of numbers are called the variance and standard deviation of the list. For a list of measurements on a scale of units like feet or inches, the SD of the list gives an indication of the typical magnitude of the difference between measurements in the list and their average, on the same scale of units as the measurements.

Example 2. Indicators.

Problem. Suppose X is the indicator of an event with probability p. Find $SD(X)$.

Solution. Since $0^2 = 0$ and $1^2 = 1$, we have $X^2 = X$. Therefore,

$$E(X^2) = E(X) = p$$

so the computational formula gives

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 = p - p^2 = p(1-p) \\ SD(X) &= \sqrt{Var(X)} = \sqrt{p(1-p)} \end{aligned}$$

Discussion. Since X has a binomial $(1, p)$ distribution, this agrees with the formula $\sqrt{npq}$ for the SD of the binomial (n, p) distribution given in Chapter 2. This formula for $n > 1$ is checked in a later example.

Example 3. Number on a die.

Problem. Let X be the number on a fair die. Find $SD(X)$.

Solution. By the computational formula

$$Var(X) = E(X^2) - \mu^2 = \frac{(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2)}{6} - (3.5)^2 = \frac{35}{12}$$

$$SD(X) = \sqrt{35/12} = 1.71$$

Scaling and Shifting

For constants a and b, $\quad SD(aX+b) = |a|SD(X)$

Shifting by a constant doesn't change the spread of the distribution, but multiplying by a or $-a$ spreads out the distribution by a factor of $|a|$. You can check this from the definition of SD, using properties of expectation. Compare with the corresponding formula for expectation:

$$E(aX+b) = aE(X)+b$$

Example 4. Celsius to Fahrenheit.

Problem. Suppose X represents a temperature in degrees Celsius, Y the same temperature in degrees Fahrenheit, so

$$Y = \frac{9}{5}X + 32$$

How are $E(Y)$ and $SD(Y)$ related to $E(X)$ and $SD(X)$?

Solution. $E(Y) = \frac{9}{5}E(X) + 32$ is $E(X)$ converted to degrees Fahrenheit. But the SD behaves differently

$$SD(Y) = \frac{9}{5}SD(X)$$

because standard deviation, as a measure of spread, is affected only by the scale factor 9/5, and not by the shift of 32.

Example 5. Successes and failures.

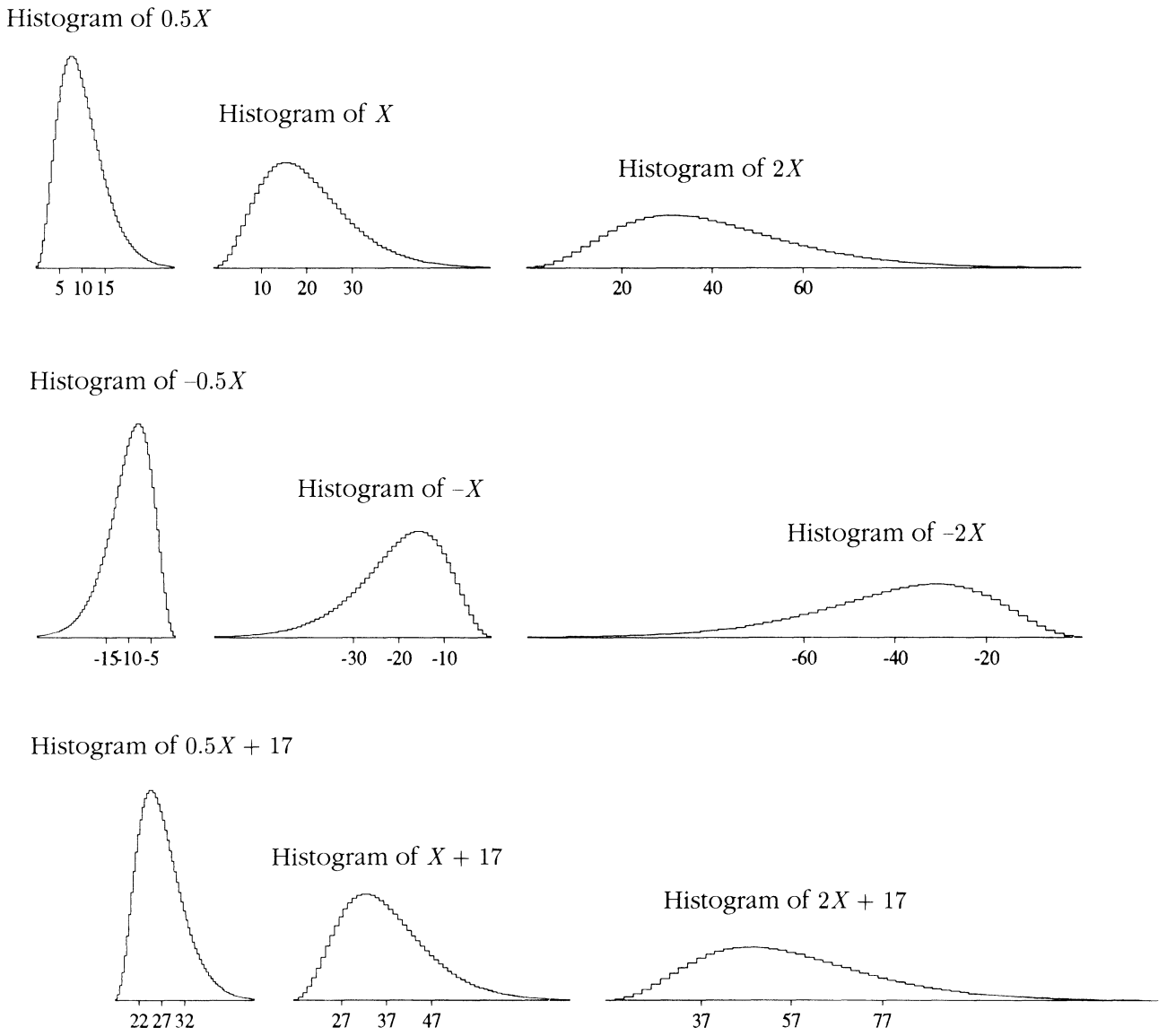
Problem. Let X be the number of successes in n trials of some kind, Y the number of failures in the same sequence of trials. Assuming that every trial results in either success or failure, how are $E(Y)$ and $SD(Y)$ related to $E(X)$ and $SD(X)$?

Solution.

$$X+Y = n \qquad \text{so} \qquad Y = n - X$$

$$E(Y) = n - E(X) \qquad SD(Y) = SD(X)$$

FIGURE 1. Scaling and shifting. The histograms display distributions of $Y = aX + b$ for various a and b. These are derived by rescaling the histogram of X shown at the center of the top row. Under the histogram of each Y are marked the points $E(Y) - SD(Y)$, $E(Y)$, and $E(Y) + SD(Y)$.



When making a normal approximation, it is convenient to transform a random variable X into a standardized variable X^*, which gives the number of SDs by which X differs from its expected value.

Standardization

If a random variable X has $E(X) = \mu$ and $SD(X) = \sigma > 0$, the random variable

$$X^* = (X - \mu)/\sigma$$

called X *in standard units,* has $E(X^*) = 0$ and $SD(X^*) = 1$.

Put another way, X^* is X relative to an origin at μ on a scale of multiples of σ. Positive values of X^* correspond to higher than expected values of X. Negative values of X^* correspond to lower than expected values of X. Any event determined by the value of X can be rewritten in terms of X^*. Usually, this is done by manipulating inequalities. For example, for any number b,

$$\begin{aligned} P(X \le b) &= P\left(\frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) \\ &= P\left(X^* \le \frac{b-\mu}{\sigma}\right) \end{aligned}$$

In case the distribution of X is approximately normal, the distribution of X^* is approximately standard normal. Then the above probability can be approximated by $\Phi[(b-\mu)/\sigma]$, where Φ is the standard normal c.d.f. For a binomial random variable X this is the normal approximation of Chapter 2, except we are now ignoring the correction from b to $b + 1/2$ (called the *continuity correction*) which is appropriate only if the range of possible values of X is a sequence of consecutive integers.

Example 6. Heights.

Problem. A person is picked at random from a population of individuals with heights distributed approximately according to the normal curve. If in this population the mean height is 5 feet 10 inches and the SD of heights is 2 inches, what approximately is the chance that the person is over 6 feet tall?

Solution. Let X represent the height of the individual. Then $E(X) = 5$ feet 10 inches and $SD(X) = 2$ inches. Converting to standard units gives

$$\begin{aligned} P(X > 6 \text{ feet}) &= P\left(\frac{X - 5 \text{ feet } 10 \text{ inches}}{2 \text{ inches}} > 1\right) \\ &= P(X^* > 1) \approx 1 - \Phi(1) \approx 16\% \end{aligned}$$

by the normal approximation.

Tail Probabilities

Consider the event that a random variable X is more than three standard deviations from its mean. To get used to some notation, look at the following six equivalent symbolic expressions of this event, in terms of

$$E(X) = \mu, SD(X) = \sigma, \text{ and } X^* = (X - \mu)/\sigma.$$

The inequalities are manipulated by adding an arbitrary constant or multiplying by a positive constant. For example, division by σ turns (1) into (6):

$$|X - \mu| > 3\sigma \tag{1}$$

$$\text{either} \quad X - \mu < -3\sigma \quad \text{or} \quad X - \mu > 3\sigma \tag{2}$$

$$\text{either} \quad X < \mu - 3\sigma \quad \text{or} \quad X > \mu + 3\sigma \tag{3}$$

$$\text{either} \quad \frac{X - \mu}{\sigma} < -3 \quad \text{or} \quad \frac{X - \mu}{\sigma} > 3 \tag{4}$$

$$\text{either} \quad X^* < -3 \quad \text{or} \quad X^* > 3 \tag{5}$$

$$|X^*| > 3 \tag{6}$$

If the distribution of X closely follows the normal curve, the probability of this event will be very small: around 3/10 of 1%, according to the normal table. But what if the distribution is not normal? How big could this probability be? 3%? or 30%? The answer is that it might be 3%, but not 30%. The largest this probability could possibly be, for any X whatsoever, is 1/9, or about 11%. This is due to the following inequality, which makes precise the idea that a random variable is unlikely to be more than a few SDs away from its mean.

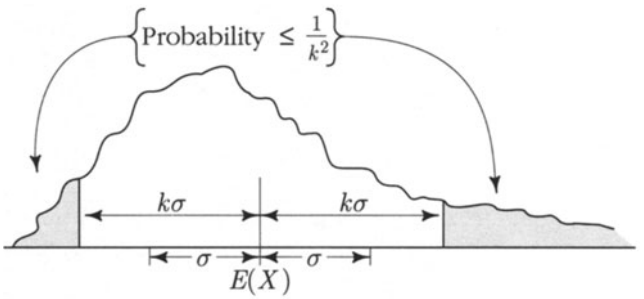
Chebychev's Inequality

For any random variable X, and any $k > 0$,

$$P[\,|X - E(X)| \geq k\,SD(X)\,] \leq \frac{1}{k^2}$$

In words: The probability that a random variable differs from its expected value by more than k standard deviations is at most $1/k^2$.

FIGURE 2. The probability bounded by Chebychev's inequality.



Proof. Let $\mu = E(X)$ and $\sigma = SD(X)$. The first step is yet another way of writing the event $[|X - \mu| \geq k\sigma]$, namely, $[(X - \mu)^2 \geq k^2\sigma^2]$. Now define $Y = (X - \mu)^2$, $a = k^2\sigma^2$, to see

$$P[|X - \mu| \geq k\sigma] = P(Y \geq a)$$

$$\leq \frac{E(Y)}{a} \quad \text{by Markov's inequality of Section 3.2, using } Y \geq 0,$$

$$= \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \quad \text{by definition of } Y, \sigma, \text{ and } a.\square$$

Comparison of the Chebychev bound with normal probabilities. Chebychev's inequality gives universal inequalities, satisfied by all distributions, no matter what their shape. For $k \leq 1$ the inequality is trivial, because then $1/k^2 \geq 1$. Here are the bounds for some values of $k \geq 1$ compared with corresponding probabilities for the normal distribution with parameters μ and σ.

Probability	Chebychev bound	Normal value
$P(\|X - \mu\| \geq \sigma)$	at most 1	0.3173
$P(\|X - \mu\| \geq 2\sigma)$	at most $1/2^2 = 0.25$	0.0465
$P(\|X - \mu\| \geq 3\sigma)$	at most $1/3^2 \approx 0.11$	0.00270
$P(\|X - \mu\| \geq 4\sigma)$	at most $1/4^2 \approx 0.06$	0.000063

As the table shows, Chebychev's bound will be very crude for a distribution that is approximately normal. Its importance is that it holds no matter what the shape of the distribution, so it gives some information about two-sided tail probabilities whenever the mean and standard deviation of a distribution can be calculated.

Example 7. Bounds for a list of numbers.

Problem. The average of a list of a million numbers is 10 and the average of the squares of the numbers is 101. Find an upper bound on how many of the entries in the list are 14 or more.

Solution. Let X represent a number picked at random from the list. Then $\mu = E(X) = 10$, $E(X^2) = 101$, so

$$\sigma = SD(X) = \sqrt{101 - 10^2} = 1,$$
$$P(X \geq 14) = P(X - \mu \geq 4\sigma) \leq P(|X - \mu| \geq 4\sigma) \leq 1/4^2,$$

by Chebychev's inequality. Consequently, the number of entries 14 or over is at most

$$10^6 P(X \geq 14) \leq 10^6/16 = 62,500$$

Remark. If the distribution of the list were known to be symmetric about 10, the probabilities $P(X \geq 14)$ and $P(X \leq 6)$ would be equal. Since it is the sum of these two probabilities which is at most $1/16$, the bound in this case could be reduced by a factor of 2 to $31,250$. If the distribution of the list were approximately normal, the number would be more like

$$10^6 \times [1 - \Phi(4)] \approx 32$$

Sums and Averages of Independent Random Variables

The main reason for the importance of variance is the following simple rule for the variance of a sum of two independent variables. This rule leads to the right SD to use in the normal approximation for a sum of n independent random variables for large n.

Addition Rule for Variances

$$Var(X + Y) = Var(X) + Var(Y) \quad \text{if} \quad X \text{ and } Y \text{ are } \textbf{independent}.$$

$$Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) \quad \text{if } X_1, \ldots, X_n \text{ are } \textbf{independent}.$$

The assumption of independence is important. In contrast to expectations, variances do not always add for dependent random variables. For example, if $X = Y$, then

$$Var(X + Y) = Var(2X) = [SD(2X)]^2 = [2\,SD(X)]^2 = 4\,Var(X)$$

while

$$Var(X) + Var(Y) = Var(X) + Var(X) = 2\,Var(X)$$

Proof of the addition rule for variances. Let $S = X + Y$. Then $E(S) = E(X) + E(Y)$, so

$$S - E(S) = [X - E(X)] + [Y - E(Y)]$$

Now square both sides and then take expectations to get

$$[S - E(S)]^2 = [X - E(X)]^2 + [Y - E(Y)]^2 + 2[X - E(X)][Y - E(Y)]$$
$$Var(S) = Var(X) + Var(Y) + 2E\{[X - E(X)][Y - E(Y)]\}$$

If X and Y are independent, then so are $X - E(X)$ and $Y - E(Y)$. So by the rule for the expectation of a product of independent variables, the last term above is the product of $E[X - E(X)]$ and $E[Y - E(Y)]$. This is zero times zero which equals zero, giving the addition rule for two independent variables. Apply this addition rule for two variables repeatedly to get the result for n variables. □

Sums of independent random variables with the same distribution. Suppose $X_1, \ldots, X_n$ are independent with the same distribution as X. You can think of the X_i as the results of repeated measurements of some kind. Because all the expectations and variances are determined by the same distribution,

$$E(X_k) = E(X) \qquad Var(X_k) = Var(X) \qquad (k = 1, \ldots, n)$$

So for the sum $S_n = X_1 + \cdots + X_n$

$$E(S_n) = nE(X) \qquad \text{by the addition rule for expectation}$$
$$Var(S_n) = nVar(X) \qquad \text{by the addition rule for variance.}$$

Taking square roots in the last formula gives the formula for $SD(S_n)$ in the next box. The results for the average follow by scaling the sum by the constant factor of $1/n$.

Square Root Law

Let S_n be the sum, $\bar{X}_n = S_n/n$ the average, of n independent random variables $X_1, \ldots, X_n$, each with the same distribution as X. Then

$$E(S_n) = nE(X) \qquad SD(S_n) = \sqrt{n}SD(X)$$

$$E(\bar{X}_n) = E(X) \qquad SD(\bar{X}_n) = \frac{SD(X)}{\sqrt{n}}$$

The expectation of a sum of n independent trials grows linearly with n. But the SD grows more slowly, according to a multiple of $\sqrt{n}$. This slow growth of the SD is due to the high probability of cancellation between terms which are above the expected value and terms which are below. The square root law for $SD(S_n)$ gives a precise mathematical measure of the extent to which this cancellation tends to occur.

Example 8. Standard deviation of the binomial distribution.

Problem. Derive the formula $\sqrt{npq}$ for the SD of the binomial (n,p) distribution.

Solution. This is the distribution of the sum $S_n = X_1 + \cdots + X_n$ of n indicators of independent events, each with probability p. So $\sqrt{npq}$ comes from the square root law for $SD(S_n)$ and the formula $\sqrt{pq}$ for the SD of an indicator, found in Example 2.

The law of averages. While as n increases $SD(S_n)$ grows as a constant times $\sqrt{n}$, dividing by n makes $SD(\bar{X}_n)$ tend to zero as a constant divided by $\sqrt{n}$. So the SD of the average of n independent trials tends to 0 as $n \to \infty$. This is an expression of the *law of averages*, which generalizes the law of large numbers stated in Section 2.2 for the proportion of successes in n Bernoulli (p) trials. Roughly speaking, the law of averages says that the average of a long sequence of independent trials $X_1, X_2, \ldots, X_n$ is likely to be close to the expected value of $X = X_1$. Here is a more precise formulation:

Law of Averages

Let $X_1, X_2, \ldots$ be a sequence of independent random variables, with the same distribution as X. Let $\mu = E(X)$ denote the common expected value of the X_i, and let

$$\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$$

be the random variable representing the average of $X_1, \ldots, X_n$. Then for every $\epsilon > 0$, no matter how small,

$$P(|\bar{X}_n - \mu| < \epsilon) \to 1 \quad \text{as} \quad n \to \infty$$

In words: as the number of variables increases, with probability approaching 1, the average will be arbitrarily close to the expected value.

Proof. From the box for the square root law, $E(\bar{X}_n) = \mu$, $SD(\bar{X}_n) = \sigma/\sqrt{n}$, where $\sigma = SD(X_1)$. Chebychev's inequality applied to $\bar{X}_n$ now gives

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P\left(|\bar{X}_n - \mu| \geq \frac{\epsilon}{SD(\bar{X}_n)} SD(\bar{X}_n)\right) \leq \left(\frac{SD(\bar{X}_n)}{\epsilon}\right)^2 = \frac{\sigma^2}{n\epsilon^2}$$

But for each fixed ϵ the right side tends to 0 as $n \to \infty$, hence so does the left side since probabilities are non-negative. Taking complements yields the result. □

Exact distribution of sums of independent variables. Suppose the X_i are independent indicator variables, with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$ for some $0 < p < 1$. For example, X_i could be the indicator of success on the ith trial in a sequence of independent trials. Then $S_n = X_1 + \cdots + X_n$ represents the number of successes in n trials, and S_n has the binomial (n, p) distribution studied in Chapter 2. In theory, and numerically by computer, the formula of Exercise 3.1.16 for the distribution of the sum of two random variables can be applied repeatedly to find the distribution of S_n for other distributions of X_i. But the resulting formulae are manageable only in a few other cases (e.g., the Poisson and geometric cases treated in the next section.)

Approximate distribution of sums of independent variables. Because there is no simple formula for the distribution of the sum S_n of n independent random variables with the same distribution as X, it is both surprising and useful that no matter what the distribution of X, there is a simple normal approximation for the distribution of S_n. This generalizes the normal approximation to the binomial distribution treated in Section 2.2.

The Normal Approximation (Central Limit Theorem)

Let $S_n = X_1 + \cdots + X_n$ be the sum of n independent random variables each with the same distribution over some finite set of values. For large n, the distribution of S_n is approximately normal, with mean $E(S_n) = n\mu$, and standard deviation $SD(S_n) = \sigma\sqrt{n}$, where $\mu = E(X_i)$ and $\sigma = SD(X_i)$. That is to say, for all $a \leq b$

$$P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

where Φ is the standard normal c.d.f. No matter what the distribution of the terms X_i, for every $a \leq b$ the error in using this normal approximation tends to zero as $n \to \infty$. The same result holds for X_i with an infinite range of possible values, provided the standard deviation is defined and finite.

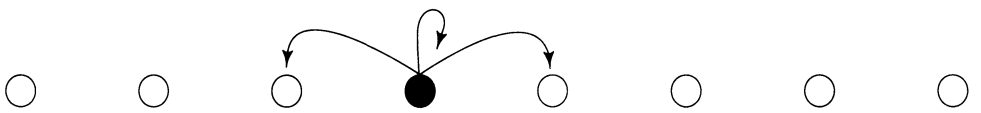
Note that the random variable $(S_n - n\mu)/\sigma\sqrt{n}$ appearing in the normal approximation is S_n in standard units. If the possible values of the X_i form a sequence of consecutive integers, the continuity correction should be used as in Section 2.2 to obtain a better approximation. The normal approximation works just as well for averages as for sums, because the factor of n has no effect on the standardized variables. For any distribution of X_i with just two possible values, the above normal approximation follows from the normal approximation to the binomial distribution, derived in Section 2.3, by using scaling properties of the mean and standard deviation to reduce to the case when the two possible values are 0 and 1. But a full proof of the central limit theorem is beyond the scope of this text.

The pictures at the end of the section show how the distribution of the sum S_n of independent and identically distributed $X_1, X_2, \ldots, X_n$ depends on the number of terms n and the common distribution of the X_i. As a general rule, the more symmetric the distribution, and the thinner its tails, the faster the approach to normality as n increases. On each page, all histograms are scaled horizontally in standard units, and vertically to keep the total area constant.

Example 9. Random walk.

Physicists use random walks to model the process of diffusion, or random motion of particles. The position S_n of a particle at time n can be thought of as a sum of displacements $X_1, \ldots, X_n$. Assuming the displacements are independent and identically distributed, the theory of this section applies.

Problem. Suppose at each step a particle moving on sites labeled by integers is equally likely to move one step to the right, one step to the left, or stay where it is.



Find approximately the probability that after 10,000 steps the particle ends up more than 100 sites to the right of its starting point.

Solution. Let X represent a single step. Then $E(X) = 0$,

$$Var(X) = E(X^2) - 0^2 = \frac{(-1)^2}{3} + \frac{0^2}{3} + \frac{1^2}{3} = \frac{2}{3}$$

and $SD(X) = \sqrt{2/3} = 0.8165$. The problem is to find $P(S_{10,000} > 100)$, where

$E(S_{10,000}) = 10,000E(X) = 0$ and

$SD(S_{10,000}) = \sqrt{10,000}\, SD(X) = 100 \times 0.8165 = 81.65$

by the square root law. The normal approximation gives

$$P(S_{10,000} > 100) = P\left(\frac{S_{10,000}}{81.65} > \frac{100}{81.65}\right) \approx 1 - \Phi(\frac{100}{81.65}) \approx 11\%$$

Skewness

Let X be a random variable with $E(X) = \mu$ and $SD(X) = \sigma$. Let $X_* = (X - \mu)/\sigma$ be X in standard units. So the first two moments of X_* are

$$E(X_*) = 0 \quad \textit{and} \quad E(X_*^2) = 1$$

The *skewness* of X, (or of the distribution of X) denoted here by Skewness(X), is the third moment of X_*:

$$\text{Skewness}(X) = E(X_*^3) = E[(X - \mu)^3]/\sigma^3$$

Skewness is a measure of the degree of asymmetry in the distribution of X. For any X with finite third moment, there is the simple formula (Exercise 33):

$$\text{Skewness}(S_n) = \text{Skewness}(X)/\sqrt{n} \tag{7}$$

for S_n the sum of n independent random variables with the same distribution as X. This implies the formula $(1 - 2p)/\sqrt{npq}$ used in Section 2.2 for the skewness of binomial (n, p) distribution.

It is easy to see that if the distribution of X is symmetric about μ, then Skewness(X) = 0. If the normal approximation to the distribution of X is good, the distribution of X must be nearly symmetric about μ, so it is be expected that Skewness(X) ≈ 0. In case Skewness(X) is significantly different from 0, the normal approximation to the distribution of X will usually not be very good. Formula (7) shows that no matter what the skewness of the distribution of X, the skewness of the sum S_n tends to zero as $n \to \infty$, though rather slowly. This is evidence of the central limit theorem: the distribution of S_n is asymptotically normal with skewness 0 in the limit, so has small skewness for large n. As in the binomial case studied in Section 2.2, an improvement to the normal approximation of S_n is obtained by replacing $\Phi(z)$ in the usual normal approximation by

$$\Phi(z) - \frac{1}{6\sqrt{n}}\text{Skewness}(X)\,(z^2 - 1)\,\phi(z)$$

where $\phi(z)$ is the standard normal curve. See Section 3.5 for an application to the Poisson distribution.

Figure 3. Distribution of the sum of n die rolls for $n = 1, 2, 4, 8, 16, 32$.

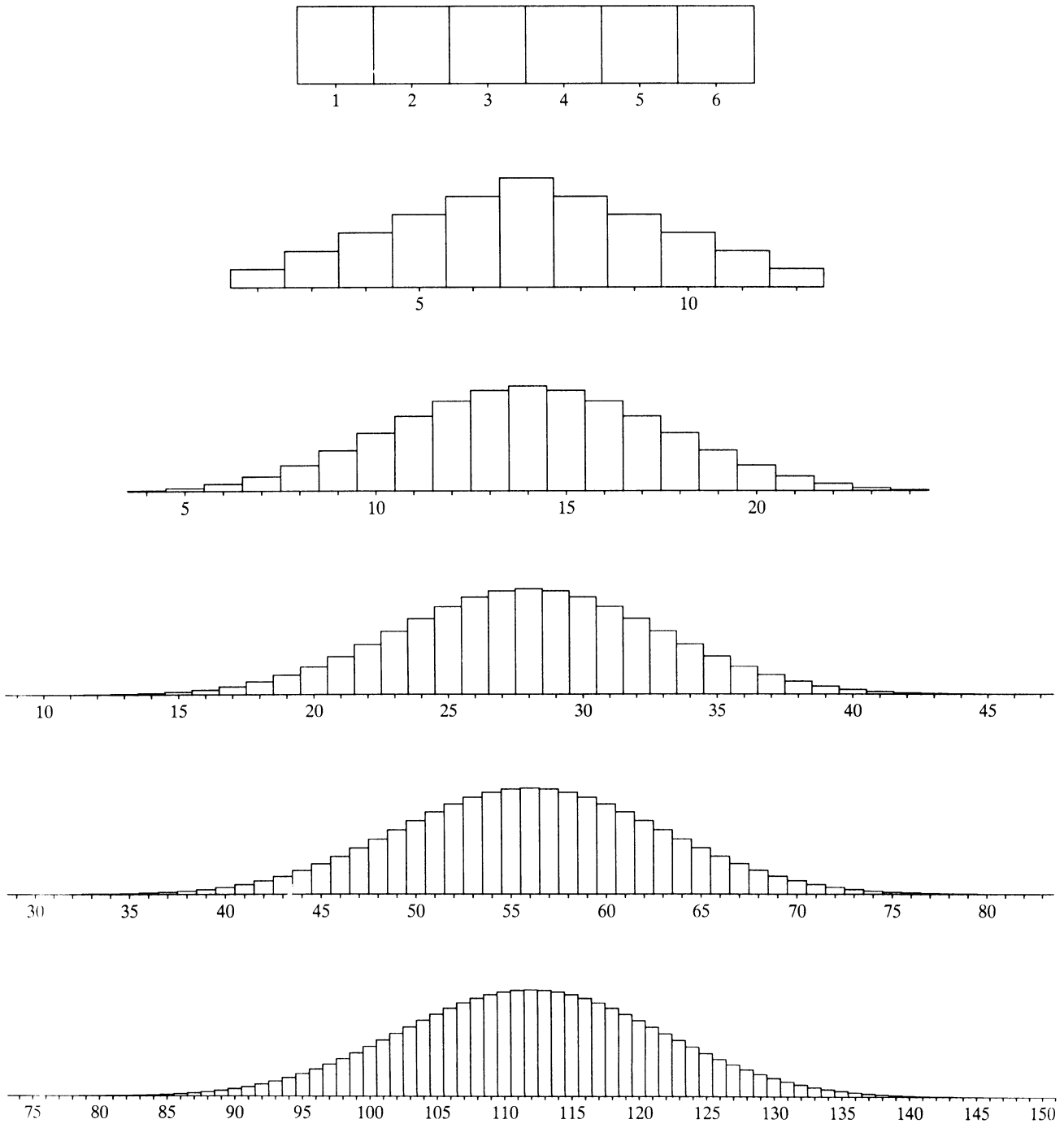


Figure 4. Distribution of S_n for $n = 1, 2, 4, 8, 16, 32$.

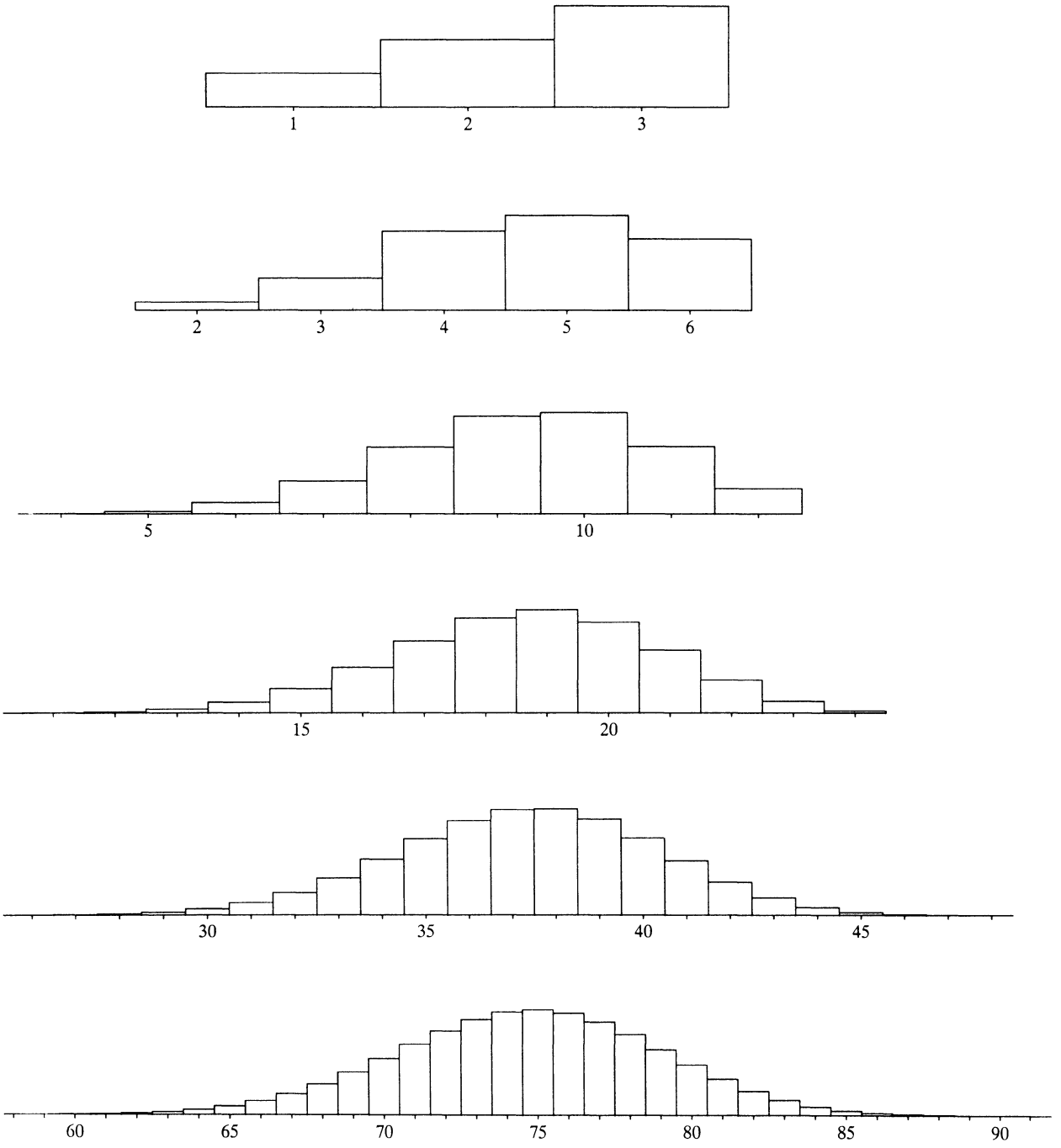
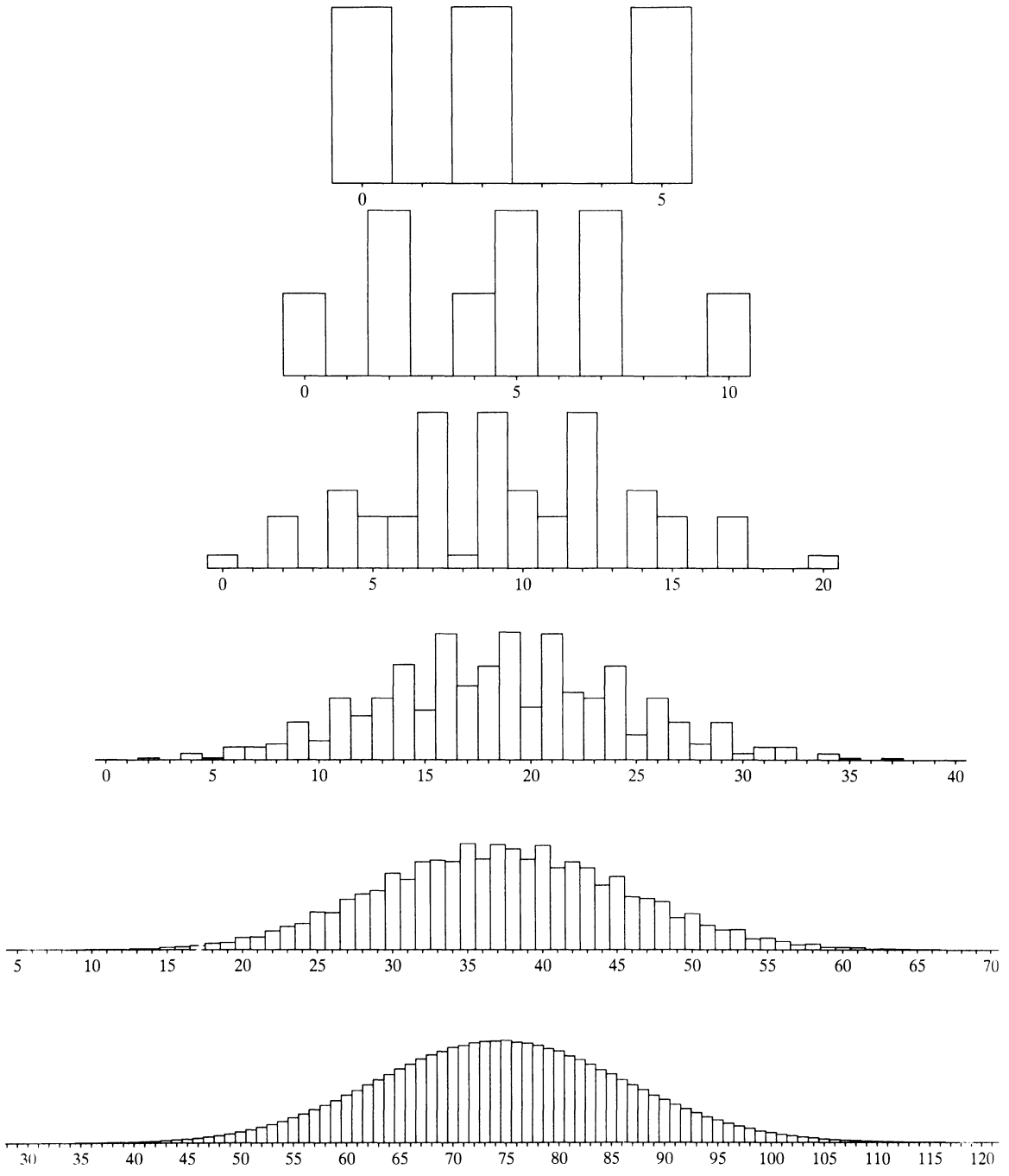


Figure 5. Distribution of S_n for $n = 1, 2, 4, 8, 16, 32$.



Exercises 3.3

1. Let X be the number of days in a month picked at random from the 12 months of a year (not a leap year).

a) Display the distribution of X in a table, and calculate $E(X)$ and $SD(X)$.

b) Repeat with X the number of days in the month containing a day picked at random from the 365 days of 1991.

2. Let Y be the number of heads obtained if a fair coin is tossed three times. Find the mean and variance of Y^2.

3. Let X, Y, and Z be independent identically distributed random variables with mean 1 and variance 2. Calculate:

a) $E(2X + 3Y)$; b) $Var(2X + 3Y)$; c) $E(XYZ)$; d) $Var(XYZ)$.

4. Suppose X_1 and X_2 are independent. Find a formula for $Var(X_1X_2)$ in terms of $\mu_1 = E(X_1)$, $\sigma_1^2 = Var(X_1)$, $\mu_2 = E(X_2)$, and $\sigma_2^2 = Var(X_2)$.

5. Show that if $E(X) = \mu$ and $Var(X) = \sigma^2$, then for every constant a

$$E\left[(X - a)^2\right] = \sigma^2 + (\mu - a)^2.$$

6. Let X_p represent the number appearing on one roll of a 'shape' which lands flat (1 or 6) with probability p, as described in Example 1.3.3. Explain without calculation why $Var(X_p)$ must increase as p increases. Then compute $Var(X_p)$ and check that it increases as p increases.

7. Suppose three marksmen shoot at a target. The ith marksman fires n_i times, hitting the target each time with probability p_i, independently of his other shots and the shots of the other marksmen. Let X be the total number of times the target is hit.

a) Is the distribution of X binomial?

b) Find $E(X)$ and $Var(X)$.

8. Let A_1, A_2, and A_3 be events with probabilities $\frac{1}{5}$, $\frac{1}{4}$, and $\frac{1}{3}$, respectively. Let N be the number of these events that occur.

a) Write down a formula for N in terms of indicators.

b) Find $E(N)$.

In each of the following cases, calculate $Var(N)$:

c) A_1, A_2, A_3 are disjoint;

d) they are independent;

e) $A_1 \subset A_2 \subset A_3$.

9. Out of n individual voters at an election, r vote Republican and $n - r$ vote Democrat. At the next election the probability of a Republican switching to vote Democrat is p_1, and of a Democrat switching is p_2. Suppose individuals behave independently. Find a) the expectation and b) the variance of the number of Republican votes at the second election.

10. Moments of the uniform distribution. Let X be uniformly distributed on $\{1, 2, \ldots, n\}$. Let $s(k, n) = 1^k + 2^k + \cdots + n^k$ be the sum of the kth powers of the first n integers.

a) Show that $E(X^k) = \dfrac{s(k,n)}{n}$ and $E\left[(X+1)^k\right] = \dfrac{s(k,n+1)-1}{n}$.

b) Deduce that $E\left[kX^{k-1} + \binom{k}{2}X^{k-2} + \cdots + 1\right] = \dfrac{(n+1)^k - 1}{n}$.

c) Use b) for $k = 2$ to obtain $E(X) = (n+1)/2$ (also obvious by symmetry), and hence $s(1, n) = n(n+1)/2$.

d) Use b) for $k = 3$ and the above formula for $E(X)$ to deduce that $E(X^2) = \frac{1}{6}(n+1)(2n+1)$ and hence $s(2, n) = \frac{1}{6}n(n+1)(2n+1)$.

e) Show that $Var(X) = (n^2 - 1)/12$.

f) Check that your formulae c) and e) agree in the case $n = 6$ with the results obtained in Example 3 for X the number on a die.

g) Use the same method to show that $s(3, n) = [s(1, n)]^2$.

[This method can be used to obtain formulae for $s(k, n)$ for an arbitrary positive integer k. But the formulae get more complicated as k increases.]

11. Suppose that Y has uniform distribution on the n numbers $\{a, a+b, \ldots, a+(n-1)b\}$, and that X has uniform distribution on $\{1, 2, \ldots, n\}$. By writing Y as a linear function of X and using results of Exercise 10, find formulae for the mean and variance of Y in terms of a, b, and n.

12. A random variable X has expectation 10 and standard deviation 5.

a) Find the smallest upper bound you can for $P(X \geq 20)$.

b) Could X be a binomial random variable?

13. Suppose the IQ scores of a million individuals have a mean of 100 and an SD of 10.

a) Without making any further assumptions about the distribution of the scores, find an upper bound on the number of scores exceeding 130.

b) Find a smaller upper bound on the number of scores exceeding 130 assuming the distribution of scores is symmetric about 100.

c) Estimate the number of scores exceeding 130 assuming that the distribution is approximately normal.

14. Suppose the average family income in an area is \$10,000.

a) Find an upper bound for the percentage of families with incomes over \$50,000.

b) Find a better upper bound if it is known that the standard deviation of incomes is \$8000.

15. a) Show that if X and Y are independent random variables, then

$$Var(X - Y) = Var(X + Y)$$

b) Let D_1 and D_2 represent two draws at random with replacement from a population, with $E(D_1) = 10$ and $SD(D_1) = 2$. Find a number c so that

$$P(|D_1 - D_2| < c) \geq 99\%$$

16. A game consists of drawing tickets with numbers on them from a box, independently with replacement. In order to play you have to stake \$2 each time you draw a ticket. Your net gain is the number on the ticket you draw. Suppose there are 4 tickets in the box with numbers $-2, -1, 0, 3$ on them. If, for example the ticket shows \$3 then you get your stake back, plus an additional \$3.

a) Let X stand for your net gain in one game. What is the distribution of X? Find $E(X)$ and $Var(X)$.

b) If you play 100 times, what is your chance of winning \$25 or more?

17. Let X be a random variable with

$$P(X = -1) = P(X = 0) = 1/4,$$

and $P(X = 1) = 1/2$. Let S be the sum of 25 independent random variables, each with the same distribution as X. Calculate approximately
a) $P(S < 0)$, b) $P(S = 0)$, and c) $P(S > 0)$.

18. In roulette, the "house special" is a bet on the five pockets 0, 00, 1, 2 and 3. There are 5 chances in 38 to win, and the bet pays 6 to 1. That is, if you place a dollar bet on the house special and the ball lands in one of the five pockets, you get your dollar back plus 6 dollars in winnings; if the ball lands in any other pocket, you lose your dollar. If you make 300 one-dollar bets on the house special, approximately what is the chance that you come out ahead?

19. A new elevator in a large hotel is designed to carry about 30 people, with a total weight of up to 5000 lbs. More than 5000 lbs. overloads the elevator. The average weight of guests at this hotel is 150 lbs., with an SD of 55 lbs. Suppose 30 of the hotel's guests get into the elevator. Assuming the weights of these guests are independent random variables, what is the chance of overloading the elevator? Give your approximate answer as a decimal.

20. Suppose you have \$100,000 to invest in stocks. If you invest \$1000 in any particular stock your profit will be \$200, \$100, \$0 or −\$100 (a loss), with probability 0.25 each. There are 100 different stocks you can choose from, and they all behave independently of each other. Consider the two cases: (1) Invest \$100,000 in one stock. (2) Invest \$1000 in each of 100 stocks.

(a) For case (1) find the probability that your profit will be \$8000 or more.

(b) Do the same for case (2).

21. Roundoff errors. Suppose you balance your checkbook by rounding amounts to the nearest dollar. Between 0 and 49 cents, drop the cents; between 50 and 99 cents, drop the cents and add a dollar. Find approximately the probability that the accumulated error in 100 transactions is greater than 5 dollars (either way)

a) assuming the numbers of cents involved are independent and uniformly distributed between 0 and 99;

b) assuming each transaction is an exact dollar amount with probability 1/4, and given not an exact dollar amount the number of cents is uniformly distributed between 1 and 99, independently for different transactions.

22. Suppose n dice are rolled.

a) Find approximately the probability that the average number is between $3\frac{5}{12}$ and $3\frac{7}{12}$ for the following values of n: 105, 420, 1680, 6720.

b) Use these values to sketch the graph of this probability as a function of n.

c) Suppose that the numbers $3\frac{5}{12}$ and $3\frac{7}{12}$ were replaced by $3\frac{1}{2} - \epsilon$ and $3\frac{1}{2} + \epsilon$ for some other small number ϵ instead of $\epsilon = \frac{1}{12}$, say $\epsilon = \frac{1}{24}$. How would this affect the graph?

23. Suppose that in a particular application requiring a single battery, the mean lifetime of the battery is 4 weeks, with a standard deviation of 1 week. The battery is replaced by a new one when it dies, and so on. Assume lifetimes of batteries are independent. What, approximately, is the probability that more than 26 replacements will have to be made in a two-year period, starting at the time of installation of a new battery, and not counting that new battery as a replacement? [*Hint*: Use the normal approximation to the distribution of the total lifetime of n batteries for a suitable n.]

24. A box contains four tickets, numbered 0, 1, 1, and 2. Let S_n be the sum of the numbers obtained from n draws at random with replacement from the box.

a) Display the distribution of S_2 in a suitable table.

b) Find $P(S_{50} = 50)$ approximately.

c) Find an exact formula for $P(S_n = k) \quad (k = 0, 1, 2, \ldots)$.

25. Equality in Chebychev's inequality. Let μ, σ, and k be three numbers, with $\sigma > 0$ and $k \geq 1$. Let X be a random variable with the following distribution:

$$P(X = x) = \begin{cases} \dfrac{1}{2k^2} & \text{if} \quad x = \mu + k\sigma \quad \text{or} \quad \mu - k\sigma \\ 1 - \dfrac{1}{k^2} & \text{if} \quad x = \mu \\ 0 & \text{otherwise.} \end{cases}$$

a) Sketch the histogram of this distribution for $\mu = 0$, $\sigma = 10$, $k = 1, 2, 3$.

b) Show that $E(X) = \mu, \quad Var(X) = \sigma^2, \quad P(|X - \mu| \geq k\sigma) = 1/k^2$.

So there is equality in Chebychev's inequality for this distribution of X. This means Chebychev's inequality cannot be improved without additional hypotheses on the distribution of X.

c) Show that if Y has $E(Y) = \mu$, $Var(Y) = \sigma^2$, and $P(|Y - \mu| < \sigma) = 0$, then Y has the same distribution as X described above for $k = 1$.

26. Mean absolute deviation.

a) Calculate the *mean absolute deviation* $E(|X - \mu|)$ for X, the number on a six-sided die.

Your answer should be slightly smaller than the standard deviation found in Example 3. This is a general phenomenon, which occurs because the operation of squaring the absolute deviations before averaging them tends to put more weight on large deviations than on small ones.

b) Use the fact that $Var(|X - \mu|) \geq 0$ to show that $SD(X) \geq E(|X - \mu|)$, with equality if and only if $|X - \mu|$ is a constant.

That is to say, unless $|X - \mu|$ is a constant, the standard deviation of a random variable is always strictly larger than the mean absolute deviation. If X is a constant, then both measures of spread are zero.

27. The SD of a bounded random variable.

a) Let X be a random variable with $0 \le X \le 1$ and $E(X) = \mu$. Show that:

(i) $0 \le \mu \le 1$; (ii) $0 \le Var(X) \le \mu(1-\mu) \le \frac{1}{4}$ [*Hint*: Use $X^2 \le X$]

b) Let X be a random variable with $a \le X \le b$ and $E(X) = \mu$. Show that:

(i) $a \le \mu \le b$; (ii) $0 \le Var(X) \le (\mu - a)(b - \mu) \le \frac{1}{4}(b-a)^2$;

(iii) $0 \le SD(X) \le (b-a)/2$.

c) The standard deviation of a list of a million digits 0, 1, 2, ..., 9 is exactly $4\frac{1}{2}$. How many nines are there in the list? Or is it impossible to answer this question without more information?

28. Let S be the number of successes in n independent Bernoulli trials, with possibly different probabilities $p_1, \ldots, p_n$ on different trials. Show that for fixed $\mu = E(S)$, $Var(S)$ is largest in case the probabilities are all equal.

29. Let $\bar{D}_n$ be the average of n independent random digits from $\{0, \ldots, 9\}$.

a) Guess the first digit of $\bar{D}_n$ so as to maximize your chance of being correct.

b) Calculate the chance that your guess is correct exactly for $n = 1, 2$, and approximately for a selection of larger values of n, and show the results in a graph.

c) How large must n be for you to be 99% sure of guessing correctly?

30. Let X_i be the last digit of D_i^2, where D_i is a random digit between 0 and 9. For instance, if $D_i = 7$ then $D_i^2 = 49$ and $X_i = 9$. Let $\bar{X}_n = (X_1 + \cdots + X_n)/n$ be the average of a large number n of such last digits, obtained from independent random digits $D_1, \ldots, D_n$.

a) Predict the value of $\bar{X}_n$ for large n.

b) Find a number ϵ such that for $n = 10,000$ the chance that your prediction is off by more than ϵ is about 1 in 200.

c) Find approximately the least value of n such that your prediction of $\bar{X}_n$ is correct to within 0.01 with probability at least 0.99.

d) Which can be predicted more accurately for large n: the value of $\bar{X}_n$, or the value of $\bar{D}_n = (D_1 + \cdots + D_n)/n$?

e) If you just had to predict the first digit of $\bar{X}_{100}$, what digit should you choose to maximize your chance of being correct, and what is that chance?

31. Normal approximation for individual probabilities. Let X be an integer valued random variable, $S_n = X_1 + \cdots + X_n$ where the X_i are independent with the same distribution as X. If the set of possible values of X contains two consecutive integers it can be shown that there is the following normal approximation to individual probabilities in the distribution of S_n:

$$P(S_n = k) \approx \frac{1}{\sqrt{2\pi n}\sigma} e^{-\frac{1}{2}(k-n\mu)^2/(n\sigma^2)} \qquad \text{where} \quad \mu = E(X) \text{ and } \sigma = SD(X)$$

This approximation holds in the sense described below formula (3) of Section 2.3, which is the special case when X has Bernoulli (p) distribution. (Note the change of notation: in formula (3), μ stands for $E(S_n)$ and σ for $SD(S_n)$.) Suppose the distribution of X is uniform on $\{0, 1, \ldots, 9\}$, as in Example 3.1.9.

a) Find μ and σ for this distribution of X.

b) Use the above normal approximation to verify the claim in the discussion of Example 3.1.9 that

$$P(S_{2m} = 9m) \sim 2/\sqrt{33\pi m} \text{ as } m \to \infty.$$

c) Let $[x]$ denote the integer part of x. Find b such that in the limit as $n \to \infty$

$$\frac{P(S_n = [(4.5)n + b\sqrt{n}\,])}{P(S_n = [(4.5)n])} \to \frac{1}{2}$$

d) For b as in part c), evaluate $\lim_{n\to\infty} P(|S_n - (4.5)n| \le b\sqrt{n}\,)$.

32. Skewness. For a random variable X with moments $\mu_k = E(X^k)$, derive the following properties of Skewnesss $(X) = E[((X - \mu)/\sigma)^3]$, where $\mu = \mu_1$ and $\sigma = \sqrt{\mu_2 - \mu^2}$ is assumed strictly positive:

a) Skewness $(X) = (\mu_3 - 3\mu\mu_2 + 2\mu^3)/\sigma^3$

b) If the distribution of X is symmetric about some point then Skewness$(X) = 0$.

c) If $a > 0$ then Skewness$(aX + b)$ = Skewness(X). What if $a < 0$?

33. Skewness of sums. Show the following:

a) If X and Y are independent with $E(X) = E(Y) = 0$ then

$$E[(X + Y)^3] = E(X^3) + E(Y^3).$$

b) If $S_n = X_1 + \cdots + X_n$ for independent X_i with the same distribution as X, then

$$\text{Skewness}(S_n) = \text{Skewness}(X)/\sqrt{n}$$

c) If S_n has binomial (n, p) distribution,

$$\text{Skewness}(S_n) = (1 - 2p)/\sqrt{npq}.$$

3.4 Discrete Distributions

Up to now, random variables were assumed to have a finite number of possible values. Probabilities and expectations were calculated as finite sums. But already in Chapter 2 useful approximations were obtained by letting the number of trials n tend to infinity. These approximations, the normal and the Poisson, lead naturally to the study of infinite outcome spaces. This section extends the basic concepts to allow a discrete distribution over an infinite sequence of possible outcomes. Important examples are the geometric and negative binomial distributions appearing in this section, and the Poisson distribution in the next. The following chapters study random variables with continuous distributions, like the uniform and normal, with an interval of possible values.

The distribution of the number of times T that you have to roll a fair die to get a six was found in Example 2 of Section 1.6:

$$P(T = i) = q^{i-1}p \qquad (i = 1, 2, \ldots)$$

where $q = 5/6$ and $p = 1/6$. This is the *geometric distribution on* $\{1, 2, 3, \ldots\}$ with parameter $p = 1/6$. Here the set of possible values of T can be counted one by one, but there is no largest possible value. This is an example of a *discrete distribution* on the positive integers.

A feature of infinite outcome spaces is that individual outcomes or sets of outcomes may be assigned probability zero. Consider, for example, the event $T = \infty$ that a six never shows up in repeated rolling of a die. This is an imaginable outcome, and you might want to include it in an outcome space. To find the probability of the event $T = \infty$ notice that if $T = \infty$, then the first n rolls are not 6. So the rules of probability imply

$$0 \le P(T = \infty) \le P(\text{first } n \text{ rolls not } 6) = (5/6)^n$$

assuming the die is fair and the rolls are independent. But since $q^n \to 0$ as $n \to \infty$ for $|q| < 1$, in particular for $q = 5/6$, this implies $P(T = \infty) = 0$.

A *discrete distribution* on the set of non-negative integers $\{0, 1, 2, \ldots\}$ is defined by a sequence of probabilities $p_0, p_1, p_2, \ldots$, such that

$$p_i \ge 0 \qquad \text{for all } i \text{ and} \qquad \sum_i p_i = 1$$

where i ranges over $0, 1, 2, \ldots$. By allowing p_i to be zero for all but a finite set of i, any distribution over a finite set labeled $0, 1, 2, \ldots, n$ could be presented like this. Probabilities involving discrete distributions can be calculated using the familiar rules of probability, together with a natural extension of the addition rule.

Infinite Sum Rule

If event A is partitioned into $A_1, A_2, A_3, \ldots$,

$$A = A_1 \cup A_2 \cup A_3 \cup \cdots \qquad \text{where} \quad A_i \cap A_j = \emptyset \qquad i \neq j$$

then

$$P(A) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

To illustrate, for a random variable X with discrete distribution on $\{0, 1, 2, \ldots\}$ given by

$$P(X = i) = p_i \qquad (i = 0, 1, \ldots)$$

$$P(X \leq 5) = \sum_{i=1}^{5} p_i$$

$$P(X > 5) = \sum_{i=6}^{\infty} p_i = 1 - \sum_{i=1}^{5} p_i$$

$$P(X \text{ is even}) = \sum_{i=0}^{\infty} p_{2i}$$

The theory of discrete distributions is mostly a straightforward extension of the theory of distributions on finite sets, treated in the previous chapters. The basic concepts of conditional probability, random variable, distribution of a random variable, joint distribution, and independence, all remain the same. All general formulae involving these concepts, in particular the rule of average conditional probabilities and Bayes' rule, remain valid simply with infinite sums of probabilities replacing finite ones. This can be proved using the infinite sum rule, which justifies familiar formulae such as

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

for discrete random variables X and Y. Here the sum over y is understood to range over the set of possible values of Y, and the infinite series can be evaluated in an arbitrary order, which is left unspecified.

Examples

Example 1. Odd or even.

Problem 1. Suppose you and I take turns at rolling a die, to see who can first roll a six. Suppose I roll first, then you roll, then I roll, and so on, until one of us has rolled a six. What is the chance that you roll the first six?

Solution. In terms of T, the number of rolls required to produce the first six, the problem is to find the probability that T is even, i.e., either 2, or 4, or 6, or .... By the infinite sum rule

$$\begin{aligned} P(T \text{ even}) &= P(T=2) + P(T=4) + P(T=6) + \cdots \\ &= qp + q^3p + q^5p + \cdots \qquad \text{where} \quad q = \frac{5}{6}, \quad p = \frac{1}{6} \\ &= qp(1 + q^2 + q^4 + \cdots) \\ &= qp/(1-q^2) \qquad \text{(geometric series with ratio } q^2\text{)} \\ &= \frac{5}{6} \times \frac{1}{6} \Big/ \left(1 - \frac{25}{36}\right) = \frac{5}{11} \end{aligned}$$

Problem 2. What is the chance that I roll the first six?

Solution. This is $P(T \text{ odd})$. Of course, a similar calculation could be done again. But there is no need. Since we argued earlier that T is certain to be finite, and then T must be either even or odd, so

$$P(T \text{ odd}) = 1 - \frac{5}{11} = \frac{6}{11}$$

Example 2. The craps principle.

Suppose A and B play over and over, independently, a game which each time results in a win for A, a win for B, or a draw (meaning no decision), with probabilities $P(A)$, $P(B)$, and $P(D)$. Suppose they keep playing until the first game that does not result in a draw, and call the winner of that game the overall winner.

Problem 1. Show that

$$P(\text{A wins overall}) = \frac{P(A)}{P(A)+P(B)} \quad \text{and} \quad P(\text{B wins overall}) = \frac{P(B)}{P(A)+P(B)}$$

Solution.

$$\begin{aligned} P(\text{A wins at game } n) &= P(\text{first } n-1 \text{ games drawn, and A wins game } n) \\ &= [P(D)]^{n-1}P(A), \qquad \text{so} \end{aligned}$$

$$P(\text{A wins}) = \sum_{n=1}^{\infty} [P(D)]^{n-1}P(A) = \frac{P(A)}{1-P(D)} = \frac{P(A)}{P(A)+P(B)}$$

Remark. Put another way, $P(\text{A wins}) = P(A \mid A \text{ or } B)$, which you may find intuitively clear without calculation. This is the basic principle behind the calculation of probabilities in the game of craps, taken up in the exercises.

Problem 2. Let G be the number of games played, X the name of the winner. Show that G has a geometric distribution, and that G and X are independent.

Solution. G is geometric with $p = 1 - P(D)$ (wait until the first nondraw)

$$\begin{aligned} P(G = n, \text{ A is the winner}) &= P(n-1 \text{ games drawn, then A wins}) \\ &= [P(D)]^{n-1} P(A) \\ &= [P(D)]^{n-1} \cdot [1 - P(D)] \cdot \frac{P(A)}{1 - P(D)} \\ &= P(G = n) \cdot P(\text{A wins}) \end{aligned}$$

Similarly, $P(G = n, \text{ B wins}) = P(G = n)P(\text{B wins})$. So G and X are independent.

Moments

The concept of expectation extends to most discrete distributions.

Expectation of a Discrete Random Variable

The *expectation* of a discrete random variable X is defined by

$$E(X) = \sum_x xP(X = x)$$

provided that the series is absolutely convergent, that is to say, provided

$$\sum_x |x| P(X = x) < \infty$$

Here X is allowed to have both positive and negative values. The assumption of absolute convergence is necessary to ensure that the value of $E(X)$ is the same, regardless of the order in which the terms are summed. If $X \geq 0$ then the expression for $E(X)$ at least always makes sense, provided that $E(X) = \infty$ is allowed as a possibility.

If $Y = g(X)$ is a numerical function of a discrete random variable X there is the usual formula

$$E[g(X)] = \sum_x g(x)P(X = x)$$

This formula holds in the sense that if either side is defined (possibly as ∞) then so is the other, and they are equal. The right side is regarded as defined provided either $g(x) \geq 0$, or the series is absolutely convergent. For example, taking X to be numerical and $g(x) = |x|$

$$E(|X|) = \sum_x |x| P(X = x)$$

This is the quantity that must be finite for $E(X)$ to be defined and finite.

Proof of these facts about expectation involves the theory of absolutely convergent series. But you need not worry about this. Just accept that the basic properties of expectation listed in Section 3.2 remain valid for discrete random variables provided finite sums are replaced where necessary by infinite ones, and it is assumed that the sums converge absolutely. It is still important to recognize a random variable as a sum of simpler ones and use the addition rule of expectation. Similar remarks apply to variance, which is defined for all random variables X with $E(X^2) < \infty$. In particular, Chebychev's inequality, the law of averages, and the normal approximation all hold for discrete random variables X with $E(X^2) < \infty$. In fact, the law of averages holds for independent and identically distributed random variables $X_1, X_2, \ldots$ provided that $E(X_1)$ is defined. But proof of this is beyond the scope of this course.

Example 3. Moments of the geometric distribution.

Let T be the waiting time until the first success in a sequence of *Bernoulli (p) trials*, meaning independent trials each of which results in either success with probability p, or failure with probability $q = 1 - p$. So T has geometric distribution on $\{1, 2, \ldots\}$ with parameter p.

Problem 1. Find $E(T)$.

Solution. $E(T) = \sum_{n=1}^{\infty} nP(T = n) = \sum_{n=1}^{\infty} nq^{n-1}p = p\Sigma_1$ where $\Sigma_1 = \sum_{n=1}^{\infty} nq^{n-1}$.

A simple formula for Σ_1 can be found by a method used also to obtain the formula for the sum Σ_0 of a geometric series

$$\Sigma_0 = 1 + q + q^2 + \cdots = 1/(1 - q)$$

Here is the calculation of Σ_1:

$$\begin{aligned} \Sigma_1 &= 1 + 2q + 3q^2 + \cdots \\ q\Sigma_1 &= \quad q + 2q^2 + \cdots \\ (1 - q)\Sigma_1 &= 1 + q + q^2 + \cdots = \Sigma_0 = 1/(1 - q) \\ \Sigma_1 &= 1/(1 - q)^2 \end{aligned}$$

This gives $E(T) = p/(1 - q)^2 = 1/p$.

Discussion. The formula $E(T) = 1/p$ is quite intuitive if you think about long-run averages. Over the long run, the average number of successes per trial is p. And the average number of trials per success is $1/p$.

Problem 2. Find $SD(T)$.

Solution. $SD(T) = \sqrt{E(T^2) - [E(T)]^2}$ where $E(T) = 1/p$ from above, and

$$E(T^2) = \sum_{n=1}^{\infty} n^2 P(T = n) = p\Sigma_2$$

where

$$\begin{aligned}
\Sigma_2 &= 1 + 4q + 9q^2 + \cdots + n^2 q^{n-1} + \cdots \\
q\Sigma_2 &= \quad q + 4q^2 + \cdots + (n-1)^2 q^{n-1} + \cdots \\
(1-q)\Sigma_2 &= 1 + 3q + 5q^2 + \cdots + (2n-1)q^{n-1} + \cdots = 2\Sigma_1 - \Sigma_0 \\
\text{so} \quad \Sigma_2 &= (1+q)/(1-q)^3
\end{aligned}$$

Substituting these expressions gives $SD(T) = \sqrt{q}/p$.

Example 4. Waiting until the rth success (negative binomial distribution).

Let T_r denote the number of trials until the rth success in Bernoulli (p) trials. To illustrate the definition, for the following sequence of results, with 1 = success, 0 = failure,

$$0001000000100100000001000000\ldots$$

$$T_1 = 4; \quad T_2 = 11; \quad T_3 = 14; \quad T_4 = 21; \quad T_5 = ??$$

Problem 1. What is the distribution of T_r?

Solution. The possible values of T_r are $r, r+1, r+2, \ldots$. For t in this range

$$\begin{aligned}
P(T_r = t) &= P(r-1 \text{ successes in first } t-1 \text{ trials, and trial } t \text{ success}) \\
&= \binom{t-1}{r-1} p^{r-1}(1-p)^{t-r} p = \binom{t-1}{r-1} p^r (1-p)^{t-r}
\end{aligned}$$

Problem 2. Find $E(T_r)$ and $SD(T_r)$.

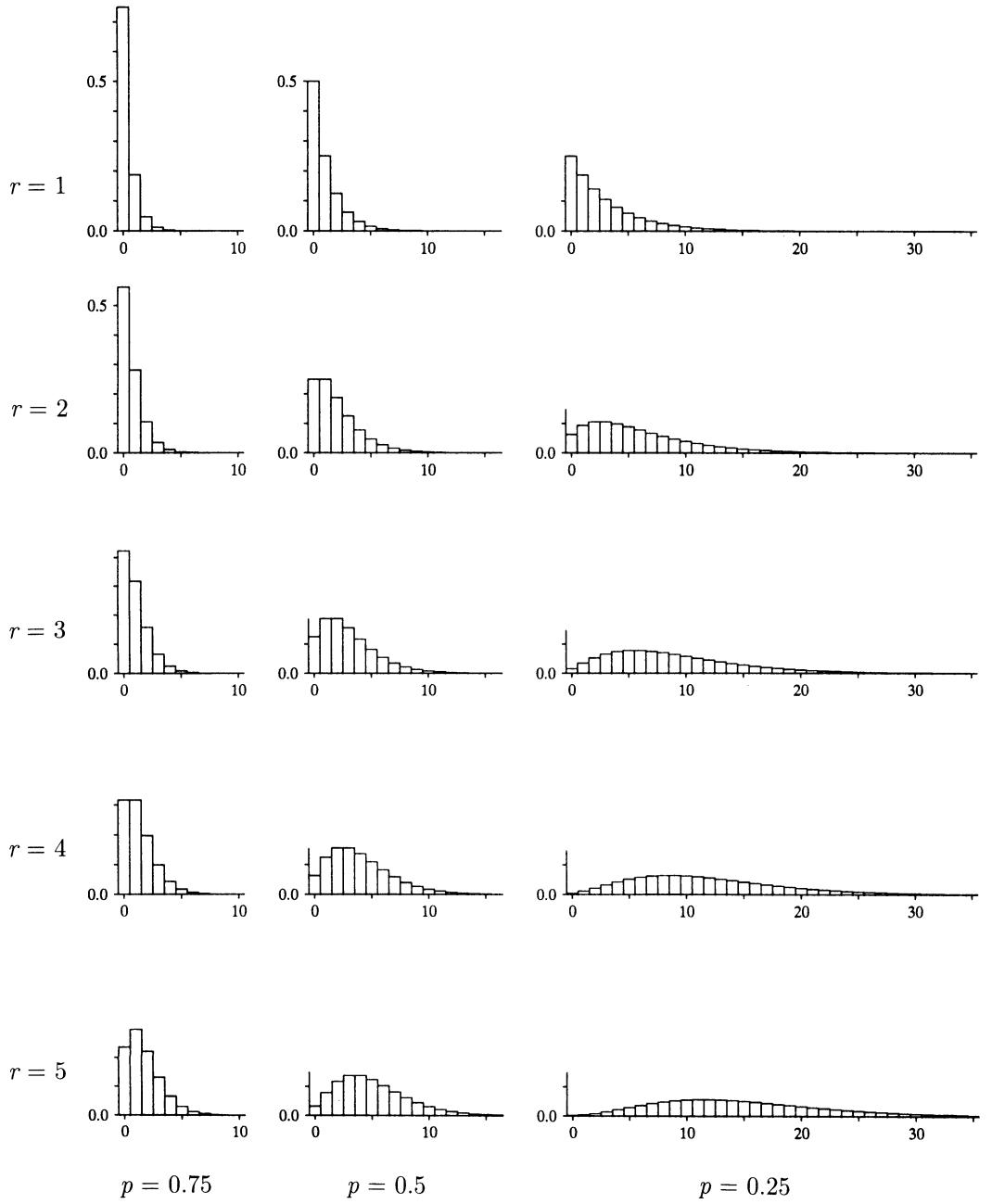
Solution. Direct calculation from the formula for the distribution is tedious. The key to a quick solution is to notice that

$$T_r = W_1 + W_2 + \cdots + W_r$$

where W_i is the waiting time after the $(i-1)$th success till the ith success. It is intuitively clear, and not hard to check, that

$$W_1, W_2, W_3, \ldots$$

FIGURE 1. **Geometric and negative binomial histograms.** The histogram in row r and column p shows the negative binomial (r, p) distribution of $T_r - r$ the number of failures before the rth success in Bernoulli (p) trials, for $r = 1, 2, 3, 4, 5$ and $p = 0.75, 0.5$, and 0.25. Note how as either p decreases or r increases, the distributions shift to the right and flatten out.



are independent, each with geometric (p) distribution. So by the results of the last example, the addition rule for expectation, and the square root law,

$$E(T_r) = r/p \qquad SD(T_r) = \sqrt{rq}/p$$

Remarks. (i) As $r \to \infty$ the distribution of T_r becomes asymptotically normal, another example of the central limit theorem. But due to the skewness of the geometric distribution of the terms being added, the approach to normality is rather slow. Particularly for p near 0.5, better approximations are obtained using the relation $P(T_r > n) = P(S_n < r)$, where S_n is the number of successes in the first n trials, and the normal approximation to the binomial (n, p) distribution of S_n.

(ii) The distribution of $T_r - r$, the number of failures before the rth success, in independent Bernoulli (p) trials, is called *negative binomial* with parameters r and p. This is just the distribution of T_r, shifted from $\{r, r+1, r+2, \ldots\}$ to $\{0, 1, 2, \ldots\}$

$$P(T_r - r = n) = P(T_r = n + r) = \binom{n+r-1}{r-1} p^r (1-p)^n \qquad (n = 0, 1, \ldots)$$

Example 5. The collector's problem.

Each box of a particular brand of cereal contains one out of a set of n different plastic animals. Suppose that the animal in each box is equally likely to be any one of the set of n, independently of what animals are in other boxes.

Problem. What is the expected number of cereal boxes a collector must buy in order to obtain the complete set of animals?

Solution. The collector gets one of the n animals in the first box. Each subsequent box contains an animal that is different from this first one with probability $(n-1)/n$, and the same with probability $1/n$. Using the independence assumption, the additional number of boxes required to get two different animals is a geometric random variable with parameter $p = (n-1)/n$ and mean

$$\frac{1}{p} = \frac{n}{n-1}$$

So the number of boxes required to get two different animals has mean

$$1 + \frac{n}{n-1}$$

Once two different animals are obtained, each box contains a new animal with probability $(n-2)/n$, and one of the old ones with probability $2/n$. So the additional

time to get three different animals once two have been obtained is a geometric random variable with parameter $p = (n-2)/n$, and mean

$$\frac{1}{p} = \frac{n}{n-2}$$

So the number of boxes required to get three different animals has mean

$$1 + \frac{n}{n-1} + \frac{n}{n-2}$$

Continuing in this way, the mean μ_n of the overall waiting time for the set of all n animals is the sum of n terms

$$\begin{aligned}\mu_n &= 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{2} + \frac{n}{1} \\ &= n\left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{2} + 1\right) \\ &= n\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)\end{aligned}$$

by reversing the order of the terms.

Discussion. To illustrate, for $n = 6$ animals, the expected number of boxes required is

$$\mu_6 = 6\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6}\right) = 14.7$$

As a variation of the problem, this is the long-run average number of times you have to roll a die in order to see every one of its faces. Similarly, the long-run average number of places you must inspect in a table of random digits, before seeing every one of the digits 0 through 9, is

$$\mu_{10} = 10\left(1 + \frac{1}{2} + \cdots + \frac{1}{10}\right) = 29.29$$

For large n, approximate values of μ_n can be obtained using Euler's approximation for the harmonic series

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \log(n) + \gamma + \frac{1}{2n}$$

where $\gamma = 0.57721\ldots$ is Euler's constant. So

$$\mu_n \approx n\log(n) + \gamma n + \frac{1}{2}$$

This approximation is good even for small n, as you can check on a calculator.

Technical remarks. The infinite sum rule looks natural enough, but there is more to it than meets the eye. Consider, for example, a sequence of mutually exclusive events A_j, each determined by a finite number of independent trials, for example, $A_j = (T = 2j)$ that a die first shows six on roll number $2j$. As j increases, so may the number of trials required to determine whether or not A_j occurs, so the event $A = A_1 \cup A_2 \cup \cdots$ may involve an unlimited number of trials, like the event $A = (T \text{ even})$ in the die example. It seems natural to *define* $P(A)$ as the sum of the infinite series

$$P(A_1) + P(A_2) + \cdots = \sum_{j=1}^{\infty} P(A_j) = \lim_{n\to\infty} \sum_{j=1}^{n} P(A_j)$$

to use three common notations for the same thing. This limit exists and is a number between 0 and 1 because the rules of probability for a finite number of trials imply that the partial sums are non-negative, increasing and bounded above by 1. That much is fairly straightforward. The hard thing to show is that this definition is consistent, because a given event A might be split up in lots of different ways, and it is not obvious that the infinite sum rule gives the same result no matter how the event A is split up. Still, mathematicians have shown that it does. So the infinite sum rule gives a consistent way of extending the definition of probability from events for a finite number of trials to events for an infinite number of trials. Mathematically, the infinite sum rule is usually taken to be an *axiom.* It is then a nontrivial theorem that the various distributions studied in this book can be defined over suitable classes of subsets so as to satisfy this axiom. Proof of this goes beyond the scope of this course; see, for example, Billingsley's book, *Probability and Measure.*

Exercises 3.4

Note: Geometric series should *not* be left unsimplified. Use

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x} \qquad (|x| < 1)$$

1. A coin which lands heads with probability p is tossed repeatedly. Assuming independence of the tosses, find formulae for

a) P(exactly 5 heads appear in the first 9 tosses);

b) P(the first head appears on the 7th toss);

c) P(the fifth head appears on the 12th toss);

d) P(the same number of heads appear in the first 8 tosses as in the next 5 tosses).

2. An urn contains 10 red balls and 10 black balls. Balls are drawn out at random with replacement until at least one ball of each color has been drawn out. Let D be the number of draws. Find: a) the distribution of D; b) $E(D)$; c) $SD(D)$.

3. Suppose you pick people at random and ask them what month of the year they were born. Let X be the number of people you have to question until you find a person who was born in December. What is $E(X)$, approximately?

4. In the game of "odd one out" three people each toss a fair coin to see if one of their coins shows a different face from the other two.

a) After one play, what is the probability of some person being the "odd one out"?

b) Suppose play continues until there is an "odd one out". What is the probability that the duration is r plays?

c) What is the expected duration of play?

5. Bill, Mary, and Tom have coins with respective probabilities p_1, p_2, p_3 of turning up heads. They toss their coins independently at the same times.

a) What is the probability it takes Mary more than n tosses to get a head?

b) What is the probability that the first person to get a head has to toss more than n times?

c) What is the probability that the first person to get a head has to toss exactly n times?

d) What is the probability that neither Bill nor Tom get a head before Mary?

6. The geometric (p) distribution on { 0, 1, 2, ... }. The geometric (p) distribution is often defined as a distribution on $\{0, 1, 2, ...\}$ instead of $\{1, 2, 3, ...\}$. A random variable W has geometric (p) distribution on $\{0, 1, 2, ...\}$ if

$$P(W = k) = q^k p \qquad (k = 0, 1, \ldots)$$

a) Show that this is the distribution of the number of failures before the first success in Bernoulli (p) trials.

b) Find $P(W > k) \qquad (k = 0, 1, \ldots)$ c) Find $E(W)$. d) Find $Var(W)$.

7. Suppose that A and B take turns in tossing a biased coin which lands heads with probability p. Suppose that A tosses first.

a) What is the probability that A tosses the first head?

b) What is the probability that B tosses the first head? For both a) and b) above, find formulae in terms of p and sketch graphs.

No matter what the value of p, A is more likely to toss the first head than B. To try to compensate for this, let A toss once, then B twice, then A once, B twice, and so on.

c) Repeat a) and b) with this scheme. Give formulae and graphs.

d) For what value of p do A and B have the same chance of tossing the first head?

e) What, approximately, is B's chance of winning for very small values of p? Give both an intuitive explanation and an evaluation of the limit as $p \to 0$ by calculus.

8. Craps. In this game a player throws two dice and observes the sum. A throw of 7 or 11 is an immediate win. A throw of 2, 3, or 12 is an immediate loss. A throw of 4, 5, 6, 8, 9, or 10 becomes the player's *point*. In order to win the game now, the player must continue to throw the dice, and obtain the point before throwing a 7. The problem is to calculate the probability of winning at craps. Let X_0 represent the first sum thrown. The basic idea of the calculation is first to calculate $P(\text{Win} | X_0 = x)$ for every possible value x of X_0, then use the law of average conditional probabilities to obtain $P(\text{Win})$.

a) Show that for $x = 4, 5, 6, 8, 9, 10$,

$$P(\text{Win} \mid X_0 = x) = P(x)/[P(x) + P(7)]$$

where $P(x) = P(X_i = x)$ is the probability of rolling a sum of x. (Refer to Example 2).

b) Write down $P(\text{Win} \mid X_0 = x)$ for the other possible values x of X_0.

c) Deduce that the probability of winning at craps is

$$P(\text{Win}) = \frac{1952}{36 \times 11 \times 10} = 0.493\ldots$$

9. Suppose we play the following game based on tosses of a fair coin. You pay me \$10, and I agree to pay you $\$n^2$ if heads comes up first on the nth toss. If we play this game repeatedly, how much money do you expect to win or lose per game over the long run?

10. Let X be the number of Bernoulli (p) trials required to produce at least one success and at least one failure. Find:

a) the distribution of X; b) $E(X)$; c) $Var(X)$.

11. Suppose that A tosses a coin which lands heads with probability p_A, and B tosses one which lands heads with probability p_B. They toss their coins simultaneously over and over again, in a competition to see who gets the first head. The one to get the first head is the winner, except that a draw results if they get their first heads together. Calculate:

a) $P(\text{A wins})$; b) $P(\text{B wins})$; c) $P(\text{draw})$;

d) the distribution of the number of times A and B must toss.

12. Let W_1 and W_2 be independent geometric random variables with parameters p_1 and p_2. Find:

a) $P(W_1 = W_2)$; b) $P(W_1 < W_2)$; c) $P(W_1 > W_2)$;

d) the distribution of $\min(W_1, W_2)$;

e) the distribution of $\max(W_1, W_2)$.

13. Consider the following gambling game for two players, Black and White. Black puts b black balls and White puts w white balls in a box. Black and White take turns at drawing at random from the box, with replacement between draws until either Black wins by drawing a black ball or White wins by drawing a white ball. Suppose Black gets to draw first.

a) Calculate $P(\text{Black wins})$ and $P(\text{White wins})$ in terms of $p = b/(b + w)$.

b) What value of p would make the game fair (equal chances of winning)?

c) Is the game ever fair?

d) What is the least total number of balls in the game, $(b + w)$, such that neither player has more than a 51% chance of winning?

14. In Bernoulli (p) trials let V_n be the number of trials required to produce either n successes or n failures, whichever comes first. Find the distribution of V_n.

15. The memoryless property. Suppose F has geometric distribution on $\{0, 1, 2, ...\}$ as in Exercise 6.

a) Show that for every $k \geq 0$,

$$P(F - k = m \,|\, F \geq k) = P(F = m), \quad m = 0, 1, ...$$

b) Show the geometric distribution is the only discrete distribution on $\{0, 1, 2, ...\}$ with this property.

c) What is the corresponding characterization of the geometric (p) distribution on $\{1, 2, \ldots\}$?

16. Fix r and p and let $P(k)$, $k = 0, 1, \ldots$, denote the probabilities in the negative binomial (r, p) distribution.

a) Show that the consecutive odds ratios are

$$P(k)/P(k-1) = (r + k - 1)q/k \qquad (k = 1, 2, \ldots)$$

b) Find a formula for the mode m of the negative binomial distribution.

c) For what values of r and p does the distribution have a double maximum? Which values k attain it?

17. Suppose the probability that a family has exactly n children is $(1 - p)p^n$, $n \geq 0$. Assuming each child is equally likely to be a boy or a girl, independently of previous children, find a formula for the probability that a family contains exactly k boys.

18. Suppose two teams play a series of games, each producing a winner and a loser, until one team has won two more games than the other. Let G be the total number of games played. Assuming your favorite team wins each game with probability p, independently of the results of all previous games, find:

a) $P(G = n)$ for $n = 2, 3, ...$;

b) $E(G)$;

c) $Var(G)$.

19. Let T_r be the number of fair coin tosses required to produce r heads. Show that:

a) $E(T_r) = 2r$;

b) $P(T_r < 2r) = 1/2$;

c) for every non-negative integer n, $\displaystyle\sum_{i=0}^{n} \binom{n+i}{n} 2^{-i} = 2^n$

20. Tail sums. Show that for a random variable X with possible values 0, 1, 2,...

a) $E(X) = \sum_{n=1}^{\infty} P(X \geq n)$;

b) $E[\frac{1}{2}X(X+1)] = \sum_{n=1}^{\infty} nP(X \geq n)$;

c) Call the first sum above Σ_1 and the second Σ_2. Find a formula for $Var(X)$ in terms of Σ_1 and Σ_2, assuming Σ_2 is finite.

21. Section 2.4 shows that the binomial (n, p) distribution approaches the Poisson (μ) distribution as $n \to \infty$, and $p \to 0$ with $np = \mu$ held fixed. Consider the negative binomial distribution with parameters r and $p = 1 - q$. Let $r \to \infty$, and let $p \to 1$ so that $rq = \mu$ is held fixed.

a) What does the mean become in the limit?

b) What does the variance become in the limit?

c) Show the distribution approaches the Poisson (μ) distribution in the limit.

22. Factorial moments and the probability generating function. The kth factorial moment of X is $f_k = E[(X)_k]$ where $(X)_k = X(X-1)\cdots(X-k+1)$. For many distributions of X with range $\{0, 1, \ldots\}$ it is easier to compute the factorial moments than the ordinary moments $\mu_k = E[X^k]$. Note that $x^n = \sum_1^n S_{n,k}(x)_k$ for some integer coefficients $S_{n,k}$. These $S_{n,k}$ are known as *Stirling numbers of the second kind.*

a) Find $S_{n,k}$ for $1 \le n \le 3$ and $1 \le k \le n$.

b) Find a formula for μ_n in terms of $f_k, 1 \le k \le n$.

c) Assuming X has non-negative integer values, let $P(X = i) = p_i$ for $i = 0, 1, \ldots$. Let $G(z) = \sum_{i=0}^{\infty} p_i z^i$, known as the *probability generating function* of X. Assume $G(r) < \infty$ for some $r > 1$. Show by switching the order of summation and differentiation k times, (which can be justified, but you need not show this) that the kth derivative $G^{(k)}(z)$ of the function $G(z)$ is $G^{(k)}(z) = \sum_{i=k}^{\infty} p_i (i)_k z^{i-k}$. Deduce that $f_k = G^{(k)}(1)$.

23. Geometric generating function and moments. Using the notation and results of Exercise 22:

a) Find the generating function of the geometric (p) distribution on $\{0, 1, 2, \ldots\}$.

b) Find the first three factorial moments of the geometric (p) distribution on the integer set $\{0, 1, 2, \ldots\}$ by differentiation of the generating function. Check the first two factorial moments yield the mean and variance as given in the text.

c) Referring to Exercise 3.3.33 for properties of skewness, use the result of b) to find the skewness of the geometric (p) distribution on $\{0, 1, 2, \ldots\}$. Without further calculation, find the skewness of the geometric (p) distribution $\{1, 2, \ldots\}$ and of the negative binomial (r, p) distribution.

24. The collector's problem. In the setting of Example 5, let T_n denote the number of boxes to get a complete set of animals.

a) Find a formula for $\sigma_n = SD(T_n)$.

b) Show that $\sigma_n < cn$ for a constant $c > 0$.

c) Deduce from Chebychev's inequality that T_n will most likely differ from $n \log n$ by only a small multiple of n.

d) (*Hard.*) Find the asymptotic distribution as $n \to \infty$ of $(T_n - n \log n)/n$. (It's not normal.)

3.5 The Poisson Distribution

The Poisson distribution is an approximation to the distribution of the number N of occurrences of events of some kind, when the events all have small probabilities, and are independent or nearly so. For example, N might be one of the following counting variables:

N_{wins}: the number of wins in n games of roulette for a gambler who bets on a single number each game.

N_{drops}: the number of raindrops which fall on a particular square inch of roof during a one-second interval of time.

$N_{\text{particles}}$: the number of radioactive particles emitted by a piece of radioactive material during an interval of time.

In case there are n independent events with equal probability p, the exact distribution of the number N that occurs is binomial (n, p). As shown in Section 2.4, if p is small this distribution is closely approximated by the Poisson distribution with parameter $\mu = E(N) = np$:

$$P(N = k) \approx e^{-\mu}\mu^k/k! \qquad (k = 0, 1, \ldots)$$

This justifies the use of the Poisson distribution in each case above. For instance, in the raindrops example, think of the square inch as divided into 100 hundredths of a square inch, each of which might or might not be hit by a raindrop. Suppose each hundredth of a square inch has the same small chance of being hit by a raindrop in the given second, independently of what happens elsewhere on the roof, and ignore the extremely small probability of the same hundredth of a square inch being hit more than once. Then N_{drops} is the number of successes in 100 independent trials, with small probability of success on each trial. You can think of $N_{\text{particles}}$ in a similar way, by dividing time into small units. By passing to a limit in which the raindrops are regarded as hitting random points in the plane, or the particles arrive at random instants on the time line, a mathematical model is obtained in which the distribution of the count is *exactly* Poisson. This is the idea of a *Poisson random scatter*, or *Poisson process*, discussed later in this section.

Features

Features of the Poisson (μ) distribution come from corresponding features of the binomial (n, p) distribution, by the passage to the limit as $n \to \infty$ and $p \to 0$ with $np = \mu$ kept fixed. It was shown in Section 2.4 that in this limit the probabilities of individual values converge

$$\binom{n}{k} p^k (1-p)^{n-k} \to e^{-\mu}\mu^k/k! \quad \text{as} \quad n \to \infty \quad \text{and} \quad p \to 0 \quad \text{with} \quad np = \mu$$

Since the binomial (n, p) distribution has mean $np = \mu$, it is natural that the Poisson (μ) limit should also have mean μ. And the SD of the binomial (n, p) distribution is $\sqrt{npq}$, which tends to $\sqrt{\mu}$ as $n \to \infty$ and $p \to 0$ with $np = \mu$.

Poisson Mean and Standard Deviation

If N has Poisson (μ) distribution,

$$E(N) = \mu \qquad SD(N) = \sqrt{\mu}$$

These formulae, made plausible by passage to the limit from binomial, will now be verified using the Poisson probability formula and the definitions of mean and SD for a discrete distribution.

Derivation of the mean.

$$\begin{aligned} E(N) &= \sum_{k=0}^{\infty} kP(N = k) \\ &= \sum_{k=1}^{\infty} ke^{-\mu}\frac{\mu^k}{k!} \\ &= e^{-\mu}\mu\sum_{k=1}^{\infty}\frac{\mu^{k-1}}{(k-1)!} \\ &= e^{-\mu}\mu\sum_{j=0}^{\infty}\frac{\mu^j}{j!} \\ &= e^{-\mu}\mu e^{\mu} = \mu \end{aligned}$$

Derivation of the SD. A direct attempt to find $E(N^2)$ would be to try to repeat the last calculation with $k^2P(N = k)$ instead of $kP(N = k)$. This gives terms of a constant times $\mu^k k^2/k!$ which are not easy to sum. But $\mu^k k(k-1)/k!$ can easily be summed, and this solves the problem:

$$\begin{aligned} E(N(N-1)) &= \sum_{k=0}^{\infty} k(k-1)e^{-\mu}\frac{\mu^k}{k!} \\ &= e^{-\mu}\sum_{k=2}^{\infty} k(k-1)\frac{\mu^k}{k!} \\ &= e^{-\mu}\mu^2\sum_{k=2}^{\infty}\frac{\mu^{k-2}}{(k-2)!} \\ &= e^{-\mu}\mu^2 e^{\mu} = \mu^2 \end{aligned}$$

$$\text{so} \quad E[N^2] = E[N(N-1)+N] = E[N(N-1)] + E(N) = \mu^2 + \mu$$
$$\text{and} \quad Var(N) = E(N^2) - [E(N)]^2 = \mu^2 + \mu - \mu^2 = \mu$$
$$SD(N) = \sqrt{\mu}$$

How μ affects the shape of the distribution. Let N_μ have Poisson (μ) distribution. For example, think of N_μ as the number of raindrops which hit a portion of a roof of area μ in a given length of time, assuming one raindrop is expected per unit area. Since N_μ has mean μ and SD $\sqrt{\mu}$, you should expect N_μ to be around μ plus or minus a small multiple of $\sqrt{\mu}$.

If μ is so close to 0 that μ^2 is negligible in comparison to μ (for example, when $\mu = 0.01$, $\mu^2 = 0.0001$), terms of order μ^2 and higher can be neglected in the expansion

$$e^{-\mu} = 1 - \mu + \mu^2/2 + \cdots$$

so

$$P(N_\mu = 0) = e^{-\mu} \approx 1 - \mu$$
$$P(N_\mu = 1) = \mu e^{-\mu} \approx \mu$$
$$P(N_\mu \geq 2) \approx 0$$

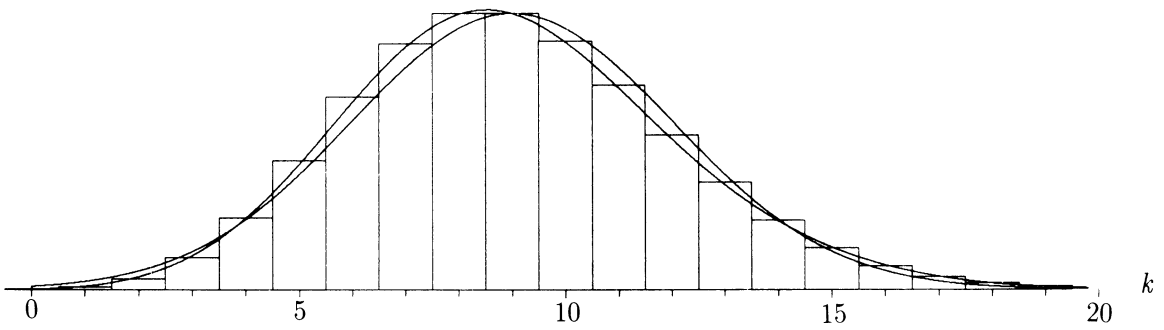
where $\approx$ means an approximation for small μ with an error of at most about μ^2. In the raindrops example, with one drop expected per unit area, this means that for a small area $\mu \ll 1$ the chance of being hit by one drop is about μ, and the chance of being hit by more than one drop is negligible in comparison.

Look again at the histograms of Poisson distributions at the end of Section 2.4. For $0 < \mu < 1$ the Poisson (μ) distribution has most probability at 0, and strictly decreasing probabilities for higher counts. As μ increases, the distribution shifts toward larger values and slowly flattens out, consistent with the formulae μ and $\sqrt{\mu}$ for the mean and SD.

Normal approximation. For μ large enough that the standard deviation $\sqrt{\mu}$ of the Poisson distribution is small in comparison to its mean μ, the distribution starts to become normal in shape. The distribution of the standardized Poisson variable $(N_\mu - \mu)/\sqrt{\mu}$ approaches standard normal as $\mu \to \infty$. This can be shown by study of consecutive odds ratios as in the binomial case treated in Section 2.3. It is yet another instance of the central limit theorem, due to the fact, discussed below, that sums of independent Poisson variables are Poisson.

Skewness. The Poisson(μ) distribution has skewness $1/\sqrt{\mu}$ (Exercise 20). Because this skewness tends to zero very slowly as $\mu \to \infty$ the approach of the Poisson distribution to normality is rather slow. Numerical calculations shown in Table 1 confirm what is apparent in Figure 1: for moderate values of μ the Poisson histogram follows a skew-normal curve much more closely than it does the normal curve.

FIGURE 1. Normal and skew-normal approximation to the Poisson (9) distribution Both the normal curve $y = \phi(z)$ and the skew–normal curve $y = \phi(z) - (1/18)\phi'''(z)$ are shown. The skew–normal curve follows the histogram much more closely.



Skew-normal Approximation to the Poisson Distribution

If N_μ has Poisson (μ) distribution, then for $b = 0, 1, \ldots$

$$P(N_\mu \le b) \approx \Phi(z) - \frac{1}{6\sqrt{\mu}}(z^2 - 1)\phi(z) \quad \text{where } z = (b + \tfrac{1}{2} - \mu)/\sqrt{\mu}.$$

Here $\Phi(z)$ is the standard normal c.d.f. and $\phi(z)$ is the standard normal curve.

It can be shown that if this skew-normal approximation is used twice to approximate interval probabilities, the worst error is less than $1/(20\mu)$ for all μ. If the skewness correction term is ignored, the resulting normal approximation with continuity but not skewness correction gives interval probabilities with much larger errors up to about $1/(10\sqrt{\mu})$ for the worst cases $a \approx \mu - \sqrt{3\mu}, b \approx \mu$ and $a \approx \mu, b \approx \mu + \sqrt{3\mu}$. If μ is sufficiently large such errors can be ignored.

The following table shows some numerical results for $\mu = 9$. The numbers are correct to three decimal places. Compare with the very similar behavior of the binomial $(100, 1/10)$ distribution displayed in Table 2 at the end of Section 2.2. As in that table, the ranges selected are the ranges over which the normal approximation is first too high, then too low, too high, and too low again. The normal approximation to the Poisson (9) distribution is very rough, but the skew-normal approximation is excellent.

TABLE 1. Approximations to the Poisson (9) distribution. The interval probability $P(a \le N_9 \le b)$ is shown for a Poisson (9) random variable N_9 along with approximations using the normal and skew–normal curves.

range of values a to b	Poisson (9) probability $P(a \le N_9 \le b)$	skew-normal approximation	normal approximation
$0-3$	0.021	0.024	0.033
$4-8$	0.434	0.431	0.400
$9-14$	0.503	0.502	0.533
$15-\infty$	0.041	0.043	0.033

Law of large numbers. Since $E(N_\mu/\mu) = \mu/\mu = 1$ and

$$SD(N_\mu/\mu) = \sqrt{\mu}/\mu = 1/\sqrt{\mu} \to 0 \quad \text{as} \quad \mu \to \infty$$

$$N_\mu/\mu \approx 1 \quad \text{for large } \mu$$

in the probabilistic sense that N_μ/μ will most likely be very close to 1. This is the law of large numbers in the Poisson context. In terms of the raindrops example, with one drop expected per unit area, this law of large numbers says that over a large area μ the average number of drops per unit area is nearly certain to be close to its expected value of 1. Both the normal approximation and the law of large numbers for the Poisson distribution are instances of more general results for sums of independent random variables, due to the result of the next paragraph.

Sums. If a big area is broken up into, say, j small areas, the number of raindrops hitting the big area is the sum of the numbers of drops in the j small areas. So the following result is very natural:

Sums of Independent Poisson Variables are Poisson

If $N_1, \ldots, N_j$ are independent Poisson random variables with parameters $\mu_1, \ldots, \mu_j$, then $N_1 + \cdots + N_j$ is a Poisson random variable with parameter $\mu_1 + \cdots + \mu_j$.

To see this via the approximation to binomial, first consider two separate blocks of Bernoulli trials of lengths n_1 and n_2 to see the following:

> If N_1 and N_2 are independent with binomial (n_1, p) and binomial (n_2, p) distributions, then $N_1 + N_2$ has binomial $(n_1 + n_2, p)$ distribution.

Now let n_1 and n_2 both tend to ∞, and $p \to 0$, with $n_1 p \to \mu_1$ and $n_2 p \to \mu_2$. Then $(n_1 + n_2)p \to \mu_1 + \mu_2$. So N_1 and N_2 approach independent Poisson variables with means μ_1 and μ_2, while $N_1 + N_2$ approaches Poisson $(\mu_1 + \mu_2)$.

Here is an alternative derivation. To simplify notation, let $\alpha = \mu_1$ and $\beta = \mu_2$.

$$\begin{aligned} P(N_1 + N_2 = k) &= \sum_{j=0}^{k} P(N_1 = j)P(N_2 = k - j) \\ &= \sum_{j=0}^{k} e^{-\alpha}\frac{\alpha^j}{j!}e^{-\beta}\frac{\beta^{k-j}}{(k-j)!} \\ &= e^{-(\alpha+\beta)}\frac{(\alpha+\beta)^k}{k!}\sum_{j=0}^{k}\frac{k!}{j!(k-j)!}\left(\frac{\alpha}{\alpha+\beta}\right)^j\left(\frac{\beta}{\alpha+\beta}\right)^{k-j} \\ &= e^{-(\alpha+\beta)}\frac{(\alpha+\beta)^k}{k!} \end{aligned}$$

because the terms in the previous sum are all the terms in a binomial distribution, with sum 1. Thus $N_1 + N_2$ has Poisson $(\alpha + \beta)$ distribution. Repeated application of this result for two terms gives the result for any number of terms.

Example 1. Number of wins.

Problem. Suppose a gambler bets ten times on events of probability 1/10, then twenty times on events of probability 1/20, then thirty times on events of probability 1/30, then forty times on events of probability 1/40. Assuming the events are independent, what is the approximate distribution of the number of times the gambler wins?

Solution. Let N_1 be the number of wins on the first 10 events of probability 1/10, N_2 the number of wins on the next 20, N_3 the number of wins on the next 30, and N_4 the number of wins on the next 40. The exact distribution of the gambler's winnings is the distribution of

$$N = N_1 + N_2 + N_3 + N_4$$

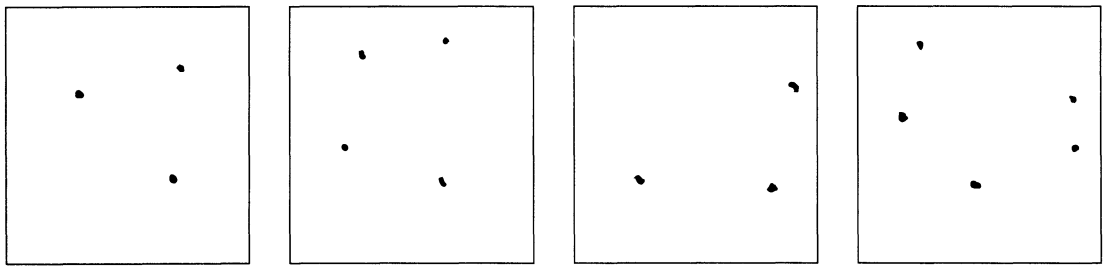
The random variables N_i, $i = 1, 2, 3, 4$, are independent, and each N_i is binomial $(10i, 1/10i)$, hence approximately Poisson (1). Thus the distribution of N must be approximately Poisson (4), by the Poisson sums theorem.

Remark. As the example suggests, the Poisson approximation to the binomial distribution extends to the case of independent trials with possibly different probabilities of success. It can be shown that if N is the number of events which occur among n independent events with probabilities $p_1, \ldots, p_n$, then provided all the probabilities p_i are small, the distribution of N is approximately Poisson (μ), where

$$\mu = E(N) = p_1 + p_2 + \cdots + p_n.$$

Random Scatter

It has already been argued informally that it would be reasonable to assume a Poisson distribution for a random variable like the number of raindrops to hit a given area in a given period of time. This idea will now be developed further to give a mathematical model for a random scatter of points in a plane such as in the diagram below.



The points might indicate, for example:

(i) points on a surface hit by particles of some kind, for example, raindrops, dust particles, atomic particles, or photons;

(ii) positions of cells of some kind on a microscopic slide;

(iii) positions of stars on a photographic plate.

The model is based on simple intuitive assumptions which turn out to imply that the number of points in a fixed area will have a Poisson distribution. The same idea of a random scatter makes sense in any number of dimensions, with length or volume instead of area. For example, a mist of raindrops is a three-dimensional scatter. And a process of random arrivals, like calls coming into a telephone exchange, can be thought of as defining a scatter of points on a time line. The basic ideas will be set out here for a scatter in two dimensions. But similar assumptions in any number of dimensions lead to the same conclusion of Poisson distributed counts.

A random scatter has both a discrete and a continuous aspect. Counting the number of points in a given region or interval gives a discrete variable. If you know enough counts for different regions you can say more or less where the points are. And the probabilities of events determined by the scatter can be derived from assumptions about the counting variables. This is the approach taken here, with assumptions which imply the counts are Poisson distributed. On the other hand, the positions in space or time of points in a scatter are typical continuous variables. Section 4.2 shows how the continuous distributions of these variables are related to the discrete Poisson distribution of counts.

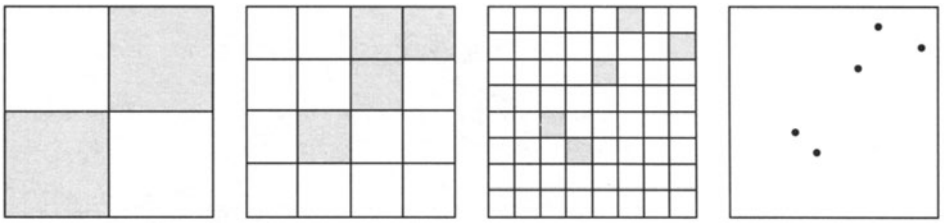
Assumptions. Consider a scatter of a finite number of points in a square. To distinguish points in the scatter from other points in the square, call the points in the

scatter *hits*. These are the places hit by the raindrops, particles or whatever, idealized as points in the square.

Assumption 1: No Multiple Hits

That is to say, distinct hits define distinct points in the square.

To state the next assumption, suppose that for each $n = 4, 16, 64, \ldots$, the square is divided into n subsquares of equal area $1/n$, as in the following diagrams. Say a subsquare is *hit* if it contains one or more hits of the scatter, and *missed* if it contains no hits. Hit squares are black and missed squares white in the diagrams. For each n, the pattern of hit squares provides some information about the scatter. This pattern gives a digital representation of the scatter, with some loss of information. As the number of subsquares n increases the pattern of hit subsquares becomes more and more sharply focused on the scatter. This can be seen in the following diagram, which shows patterns derived from a scatter of 5 points in the square.



Assumption 2: Randomness of Hits on Subsquares

For each n, any one of the n subsquares is hit with the same probability, say p_n, independently of hits on the other $n - 1$ subsquares.

Note that the randomness assumption refers separately to each digital representation. The digital representations of a random scatter for different values of n are, in fact, highly dependent. If you know the digital representation for some value of n, the representation for smaller values of n is completely determined.

Poisson Scatter Theorem

The assumptions of no multiple hits and randomness imply there is a positive constant λ such that:

(i) for each subset B of the square, the number $N(B)$ of hits in B is a Poisson random variable with mean $\lambda \times \text{area}(B)$;

(ii) for disjoint subsets $B_1, \ldots, B_j$, the numbers of hits $N(B_1), \ldots, N(B_j)$ are mutually independent.

The random scatter is then called a *Poisson scatter with intensity* λ. The intensity is the expected number of hits per unit area. Conversely, (i) and (ii) imply the assumptions of no multiple hits and randomness.

A proof of the Poisson scatter theorem is sketched at the end of the section.

Global interpretation of the intensity λ. If the scatter in the square is just part of a Poisson scatter over a larger area, the law of large numbers shows that

λ is the limiting average number of hits per unit area over a large area.

Local interpretation of the intensity λ. This refers to sets B with small area. From the Poisson distribution of $N(B)$,

$$P(\text{one hit on } B) = \lambda \,\text{area}(B)\, e^{-\lambda\, \text{area}(B)} \sim \lambda\, \text{area}(B) \quad \text{as} \quad \text{area}(B) \to 0$$

and the probability of two or more hits on B is negligible in comparison. So

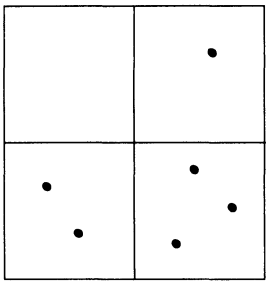
λ is the probability of a hit per unit area, as the area tends to zero.

Sums again. The fact that sums of independent Poisson variables are again Poisson is built into the concept of a Poisson scatter. For if $B_1, \ldots, B_j$ is a partition of a unit square into sets with areas $p_1, \ldots, p_j$, where $\sum_i p_i = 1$, then the total number of hits is $N = \sum_i N(B_i)$. If the scatter is Poisson with intensity λ, then N is Poisson (λ), while the $N(B_i)$, $1 \le i \le j$, are independent Poisson variables with means λp_i, which could be any positive numbers with sum λ.

Scatters over other sets. The theorem extends to scatters over other subsets of the plane than a square, and scatters on the line or in higher dimensions. Then length or volume replaces area.

Example 2. Particle hits.

Problem. Suppose particles hit a square at random according to a Poisson random scatter, with 8 particles expected in the whole square. What is the probability that the four equal subsquares in the diagram are hit by exactly 0, 1, 2, and 3 particles, respectively?



Solution. Since the numbers of hits on the four squares are independent Poisson random variables, all with parameter $8 \times 1/4 = 2$, the probability in question is

$$\frac{e^{-2}2^0}{0!} \times \frac{e^{-2}2^1}{1!} \times \frac{e^{-2}2^2}{2!} \times \frac{e^{-2}2^3}{3!} = \frac{e^{-8}2^6}{12}$$

Example 3. Bacterial colonies.

Problem 1. Suppose a volume of 1000 drops of water contains 2000 bacteria, separate from each other and thoroughly mixed in the water. A single drop is smeared uniformly over the surface of a dish. The dish contains nutrients on which the bacteria feed and multiply. After a few days, wherever a bacterium was deposited on the dish a visible colony of bacteria appears. Find the distribution of the number of colonies that appear: a) over the whole plate, b) over an area of half the plate.

Solution. It seems reasonable to suppose that the positions of bacterial colonies over the plate form a Poisson random scatter. Since 1000 drops contain 2000 bacteria, the expected number of bacteria per drop may be estimated as $2000/1000 = 2$. So the distribution of the number of bacteria on the whole plate is Poisson with mean 2. And the distribution of the number in half the plate is Poisson with mean 1.

Remark. Instead of thinking of the scatter over the plate to justify the Poisson distribution, you might think that each of the 2000 bacteria was present in the drop smeared on the plate with probability 1/1000, independently of the others. Then the number of bacteria on the plate would have binomial $(2000, 1/1000)$ distribution, which is Poisson (2) for all practical purposes. Similarly, for the number on half the plate, you get binomial $(2000, 1/2000)$, which is approximately Poisson (1). But the assumption of random scatter implies that the numbers in the two halves of the plate are independent, something not so obvious by the second method.

Problem 2. Suppose now it is not certain that a bacterium will survive and produce a visible colony, but that this happens with probability p for each bacterium on the plate, independently of the others. What now is the distribution for the number of colonies?

Solution. It is intuitively clear that the scatter of colonies must still satisfy the hypotheses of a Poisson scatter. The intensity of the colonies can be calculated from its local interpretation. Take the area of the whole plate to be 1, so by the previous example the intensity for the scatter of all bacteria landing on the plate is 2 per unit area, and

take a region B so small that

$$P(\text{one bacterium in } B) \approx 2\,\text{area}(B) \qquad P(2 \text{ or more bacteria in } B) \approx 0$$

where $\approx$ allows an error of order area(B) squared. Then

$$\begin{aligned} P(\text{one colony in } B) &= P(\text{one bacterium in } B \text{ and colony}) \\ &\quad + P(2 \text{ or more bacteria in } B \text{ and colony}) \\ &\approx 2\,\text{area}(B)\,p = 2p\,\text{area}(B) \end{aligned}$$

So the scatter of colonies has intensity $2p$ per unit area. The number of colonies on the whole plate therefore has Poisson $(2p)$ distribution.

Remark. Again, the same conclusion can be obtained another way. Think of the number of colonies as the sum of 2000 independent indicator random variables, indicating whether or not each of the 2000 bacteria gets deposited on the plate and then produces a colony. The chance of a bacterium getting on to the plate is $1/1000$, and the chance of it producing a colony, given that it gets on the plate, is p. So the overall probability of being deposited on the plate and then surviving is $p/1000$. This makes the number of colonies have binomial $(2000, p/1000)$ distribution, which is Poisson $(2p)$ for all practical purposes.

The last example illustrates a useful property of Poisson scatters, which can be derived in general by the same argument:

Thinning a Poisson Scatter

Suppose that in a Poisson scatter with intensity λ, each point of the scatter is kept with probability p, and erased (or *thinned*) with probability $1 - p$, independently both of the positions of points in the scatter and of all other thinnings. Then the scatter of points that are kept is a Poisson scatter with intensity λp.

Similarly, the scatter of points that are thinned is a Poisson process with intensity λq, where $q = 1 - p$. It can be shown, moreover, that the two scatters, one of points that are kept, and the other of points that are thinned, are independent. This means that any event determined by the numbers and positions of points in one scatter is independent of any such event determined by the other. In the example with the bacterial colonies, the numbers and positions on the plate of the bacteria that survive to produce colonies are independent of the numbers and positions of those that do not.

If you combine or *superpose* these two independent Poisson scatters, with intensities, say, $\alpha = \lambda p$ and $\beta = \lambda q$, you get back the original Poisson scatter with intensity $\lambda = \alpha + \beta$. So thinning can be understood as a kind of inverse to the more obvious

operation of superposition of two independent Poisson scatters, which gives a new Poisson scatter whose intensity is the sum of the intensities of the component scatters.

Sketch Proof of the Poisson Scatter Theorem

Step 1. Poisson distribution for the total number of hits. Let N be the total number of hits in the whole square, assumed to be of unit area. Let N_n be the number of subsquares hit when the unit square is divided into n subsquares. Then N_n increases as n increases, because each hit on one of the n subsquares must contribute one or more hits to all counts with more subsquares. And $N_n = N$ for all n large enough that the distance across one of the n subsquares is shorter than the smallest distance between two of the hits in the scatter, since then the N hits must fall in N different subsquares. (This is where the assumption of no multiple hits is essential.) Just how large n must be before $N_n = N$ depends on the scatter. But whatever the scatter, N_n eventually equals N. So the distribution of N can be found as the limit as $n \to \infty$ of the distribution of N_n. (Technically, this uses the infinite sum rule for probabilities, taken here as an axiom.) By the randomness assumption, N_n has binomial (n, p_n) distribution, where p_n is the probability that one of the subsquares of area $1/n$ is occupied. Since N_n increases with n, so does its expectation np_n. Therefore np_n converges to a limit λ as $n \to \infty$, and you can show that λ must be finite (exercise). Consequently, the limit distribution of N_n is Poisson (λ). This is the distribution of N.

Step 2. Poisson distribution for the number of hits on a subset B. Assuming B is a *simple* subset of the unit square, meaning a finite union of subsquares at some level, this is similar to the argument above, with N replaced by $N(B)$ and N_n replaced by $N_n(B)$, the number of hit squares of area $1/n$ within B. For large enough n, the simple set B is the union of some number n_B of subsquares of area $1/n$. In fact, $n_B = n\,\text{area}(B)$, since we assume the whole square has unit area, so $\text{area}(B) = n_B/n$. Now $N_n(B)$ has binomial (n_B, p_n) distribution, where

$$n_B p_n = np_n\,\text{area}(B) \to \lambda\,\text{area}(B) \quad \text{as} \quad n \to \infty$$

So in the limit the distribution of $N(B)$ is Poisson with mean $\lambda\,\text{area}(B)$. The same conclusion for more general subsets B is justified by approximation arguments or measure theory.

Step 3. Independence of counts in disjoint subsets. This comes from the assumed independence of hits in different subsquares, by letting the number of subsquares tend to infinity. □

Exercises 3.5

1. Suppose 1% of people in a large population are over 6 feet 3 inches tall. Approximately what is the chance that from a group of 200 people picked at random from this population, at least four people will be over 6 feet 3 inches tall?

2. How many raisins must cookies contain on average for the chance of a cookie containing at least one raisin to be at least 99%?

3. The cookie dough used by a bakery to make 2-ounce cookies contains an average of 32 raisins per pound of dough. The bakery sells cookies in bags of a dozen.

a) Suppose that customers complain if one or more of the cookies in a bag contains no raisins. Over the long run, about what proportion of bags of cookies give rise to complaints?

b) Approximately what average number of raisins per pound would ensure that only 5% of the bags give rise to complaints?

4. Books from a certain publisher contain an average of 1 misprint per page. What is the probability that on at least one page in a 300-page book from this publisher there will be at least 5 misprints?

5. Microbes are smeared over a plate at an average density of 5000 per square inch. The viewing field of a microscope is 10^{-4} square inches of this plate. What is the chance that at least one microbe is in the viewing field? What assumptions are you making?

6. Suppose rain is falling at an average rate of 30 drops per square inch per minute. What is the chance that a particular square inch is not hit by any drops during a given 10-second period? What assumptions are you making?

7. Suppose raisin muffins from the recycling bakery have an average of 3 fresh raisins and 2 rotten raisins per muffin.

a) What is an appropriate distribution for the number of each kind of raisin, and for the total?

b) If you bite off 20% of a muffin, what is the probability you get no raisins?

8. A Geiger counter receives pulses at an average rate of 10 per minute. What is the probability of three pulses appearing in a given half-minute period? What assumptions are you making?

9. Suppose that X and Y are independent Poisson random variables with parameters 1 and 2, respectively. Find:

a) $P(X = 1 \text{ and } Y = 2)$;

b) $P\left(\frac{X+Y}{2} \geq 1\right)$;

c) $P\left(X = 1 | \frac{X+Y}{2} = 2\right)$

10. Let X have Poisson (λ) distribution. Calculate:

a) $E(3X + 5)$; b) $Var(3X + 5)$; c) $E\left[\frac{1}{1+X}\right]$.

11. Suppose X, Y, and Z are independent Poisson random variables, each with mean 1. Find

a) $P(X + Y = 4)$; b) $E\left[(X + Y)^2\right]$; c) $P(X + Y + Z = 4)$.

12. Radioactive substances emit α-particles. The number of such particles reaching a counter over a given time period follows the Poisson distribution. Suppose two substances emit α-particles independently of each other. The first substance gives out α-particles which reach the counter according to the Poisson (3.87) distribution, while the second substance emits α-particles which reach the counter according to the Poisson (5.41) distribution. Find the chance that the counter is hit by at most 4 particles.

13. Regard the positions of molecules in a room as the points of a Poisson random scatter in 3 dimensions. According to physics, there are about 6.023×10^{23} molecules in every 22.4 liters of air at normal temperature and pressure. (A liter is 1000 cubic centimeters.) Let $N(x)$ be the random number of molecules in a particular cube of air with sides of length x centimeters.

a) Calculate the mean $\mu(x)$ and standard deviation $\sigma(x)$ of $N(x)$.

b) How small does x have to be in order that $\sigma(x)$ be 1% of $\mu(x)$, so fluctuations in density of around 1% over a cube of length x are likely to occur?

14. Assume that each of 2000 individuals living near a nuclear power plant is exposed to particles of a certain kind of radiation at an average rate of one per week. Suppose that each hit by a particle is harmless with probability $1 - 10^{-5}$, and produces a tumor with probability 10^{-5}. Find the approximate distribution of:

a) the total number of tumors produced in the whole population over a one-year period by this kind of radiation;

b) the total number of individuals acquiring at least one tumor over a year from this radiation.

Sketch the histograms of each distribution, and find the means and SD's.

15. A book has 200 pages. The number of mistakes on each page is a Poisson random variable with mean 0.01, and is independent of the number of mistakes on all other pages.

(a) What is the expected number of pages with no mistakes? What is the variance of the number of pages with no mistakes?

(b) A person proofreading the book finds a given mistake with probability 0.9. What is the expected number of pages where this person will find a mistake?

(c) What, approximately, is the probability that the book has two or more pages with mistakes?

16. On average, one cubic inch of Granma's cookie dough contains 2 chocolate chips and 1 marshmallow.

a) Granma makes a cookie using three cubic inches of her dough. Find the chance that the cookie contains at most four chocolate chips. State your assumptions.

b) Assume the number of marshmallows in Granma's dough is independent of the number of chocolate chips. I take three cookies, one of which is made with two cubic inches of dough, the other two with three cubic inches each. What is the chance that at most 1 of my cookies contains neither chocolate chips nor marshmallows?

17. Raindrops are falling at an average rate of 30 drops per square inch per minute.

a) What is the chance that a particular square inch is not hit by any drops during a given 10-second period?

b) If each drop is a big drop with probability $2/3$ and a small drop with probability $1/3$, independently of the other drops, what is the chance that during 10 seconds a particular square inch gets hit by precisely four big drops and five small ones?

18. A population comprises X_n individuals at time $n = 0, 1, 2, \ldots$. Suppose that X_0 has Poisson (μ) distribution. Between time n and time $n+1$ each of the X_n individuals dies with probability p, independently of the others. The population at time $n+1$ is formed from the survivors together with a random number of immigrants who arrive independently according to a Poisson (μ) distribution.

a) What is the distribution of X_n?

b) What happens to this distribution as $n \to \infty$?

19. Poisson generating function and moments. Suppose X has Poisson(μ) distribution. Using the notation and results of Exercise 3.4.22,

a) Show that $G(z) = e^{-\mu+\mu z}$.

b) Find the first three factorial moments X.

c) Deduce the values of the first three ordinary moments of X.

d) Show that $E(X-\mu)^3 = \mu$ and Skewness$(X) = 1/\sqrt{\mu}$.

20. Skewness of the Poisson(μ) distribution. Derive the formula $1/\sqrt{\mu}$ for the skewness of the Poisson(μ) distribution from the Poisson approximation to binomial distribution (you can assume the required switches of sums and limits are justified).

21. Skew-normal approximation to the Poisson distribution. Derive the skew-normal approximation to the Poisson (μ) distribution stated in this section:

a) from the skew-normal approximation to the binomial (n, p) distribution (in Section 2.2) by passage to the Poisson limit as $n \to \infty$ and $p \to 0$ with $np = \mu$;

b) from the skew-normal approximation for the sum of n independent random variables stated at the end of Section 3.3.

c) For N_{10} with Poisson (10) distribution, find $P(N_{10} \le 10)$ correct to three significant figures.

d) Find the normal approximation to $P(N_{10} \le 10)$ with continuity but not skewness correction, "correct" to three significant figures. Observe that the last two figures are useless: the error of approximation exceeds 0.02.

e) Find the normal approximation to $P(N_{10} \le 10)$ with continuity and skewness correction, correct to three significant figures. [All three figures should be correct. The actual error of approximation is about 2×10^{-5}.]

3.6 Symmetry (Optional)

This section studies a symmetry property for joint distributions, and illustrates it by applications to sampling without replacement. Let (X, Y) be a pair of random variables with joint distribution defined by

$$P(x, y) = P(X = x, Y = y)$$

The joint distribution is called *symmetric* if $P(x, y)$ is a symmetric function of x and y. That is to say,

$$P(x, y) = P(y, x) \qquad \text{for all } (x, y)$$

Graphically, this means that the distribution in the plane is symmetric with respect to a flip about the upward sloping diagonal line $y = x$. A glance at the figure on page 148 shows that a symmetric joint distribution is obtained for X and Y derived by sampling either with or without replacement from the set $\{1, 2, 3\}$. A symmetric joint distribution is obtained more generally whenever X and Y are two values picked by random sampling from some arbitrary list of values, either with or without replacement. This is obvious for sampling with replacement, and verified below for sampling without replacement.

In terms of random variables, the joint distribution of (X, Y) is symmetric if and only if (X, Y) has the same joint distribution as (Y, X). Then X and Y are called *exchangeable*. If X and Y are exchangeable then X and Y have the same distribution. This is true by the change of variable principle: X is a function (the first coordinate) of (X, Y), and Y is the same function of (Y, X).

The joint distribution of three random variables X, Y, and Z is called *symmetric* if

$$P(x, y, z) = P(X = x, Y = y, Z = z)$$

is a symmetric function of (x, y, z). That is to say, for all (x, y, z)

$$P(x, y, z) = P(x, z, y) = P(y, x, z) = P(y, z, x) = P(z, x, y) = P(z, y, x)$$

(all 3! = 6 possible orders of x,y and z). Equivalently, the 6 possible orderings of the random variables,

$$(X, Y, Z),\ (X, Z, Y),\ (Y, X, Z),\ (Y, Z, X),\ (Z, X, Y),\ (Z, Y, X)$$

all have the same joint distribution. Then X, Y, and Z have the same distribution, and each of the three pairs (X, Y), (X, Z), and (Y, Z) has the same (exchangeable) joint distribution, by the change of variable principle again.

A function of n variables, say $f(x_1, \ldots, x_n)$, is called *symmetric* if the value of f remains unchanged for all of the $n!$ possible permutations of the variables. Examples of symmetric functions are the sum $g(x_1) + g(x_2) + \cdots + g(x_n)$ and the product $g(x_1)g(x_2)\cdots g(x_n)$ for any numerical function $g(x)$.

Symmetry of a Joint Distribution

Let $X_1, \ldots, X_n$ be random variables with joint distribution defined by

$$P(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

The joint distribution is *symmetric* if $P(x_1, \ldots, x_n)$ is a symmetric function of $(x_1, \ldots, x_n)$. Equivalently, all $n!$ possible orderings of the random variables $X_1, \ldots, X_n$ have the same joint distribution. Then $X_1, \ldots, X_n$ are called *exchangeable*. Exchangeable random variables have the same distribution. For $2 \leq m \leq n$, every subset of m out of n exchangeable random variables has the same symmetric joint distribution of m variables.

The simplest example of an exchangeable sequence of random variables is n independent trials $X_1, \ldots, X_n$. Then

$$P(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2) \cdots p(x_n)$$

where $p(x) = P(X_i = x)$ defines the common distribution of the X_i. This a symmetric function of $(x_1, x_2, \ldots, x_n)$ because the product is the same evaluated in any order. Sampling with replacement is a special case of independent trials. Here is a more interesting example:

Sampling Without Replacement

The basic setup for sampling without replacement was described in Section 2.5. Suppose there is some population of N individuals. Suppose the ith individual in the population has some attribute b_i, for example the color of the ith ball in a box, or the height of the ith individual in a human population. Suppose n items are drawn one by one without replacement from the population. Let X_j be the attribute of the jth individual in the sample. So $X_1, \ldots, X_n$ might represent the random sequence of colors of n balls drawn at random without replacement from a box, or the random sequence of heights in a sample without replacement from a human population.

Symmetry in Sampling Without Replacement

Let $X_1, \ldots, X_n$ be a sample of size n without replacement from a list of values $\{b_1, \ldots, b_N\}$, where $2 \leq n \leq N$. Then $X_1, \ldots, X_n$ are exchangeable. In particular, for $1 \leq m \leq n$ the joint distribution of any subset of m of the X_i has the same distribution as a random sample of size m without replacement from the list $\{b_1, \ldots, b_N\}$.

This is proved in three stages as follows:

Proof for $n = N$ and $b_i = i$, $1 \le i \le n$. In this case $(X_1, \ldots, X_n)$ is an exhaustive random sample without replacement from the list $1, 2, \ldots n$, that is, a random permutation of $\{1, 2, \ldots, n\}$, as in Example 3.1.6. The joint probability function was calculated in that example and found to be symmetric. So $(X_1, \ldots, X_n)$ is exchangeable. □

Remark. The exchangeability of a random permutation is quite intuitive if you think of generating $(X_1, \ldots, X_n)$ by shuffling and then dealing out in order a deck of n cards labeled $1, 2, \ldots, n$. Any particular rearrangement of the variables $(X_1, \ldots, X_n)$ then corresponds to a particular deterministic shuffle before the deal. And it is intuitively clear that any particular additional deterministic shuffle of a perfectly shuffled deck must keep the deck perfectly shuffled. The exchangeability of a random permutation $(X_1, \ldots, X_n)$ is not so intuitive, but still true, for $X_1, \ldots, X_n$ generated by drawing balls at random one by one from an urn containing n balls labeled $1, 2, \ldots, n$.

Proof for $n = N$ and a general list $\{b_1, \ldots, b_n\}$. Now $\{b_1, \ldots, b_n\}$ can be any list of values whatever, allowing repetitions of values. The values need not be numerical. For example, for $n = N = 6$, $b_1 = b_2 = b_3 = b$, $b_4 = b_5 = r$, and $b_6 = w$, might represent a listing of the colors of balls in a box of 3 black balls, 2 red balls, and 1 white ball. A typical result of 6 draws from the box without replacement would then be the event

$$(X_1, X_2, X_3, X_4, X_5) = (b, r, w, b, b, r)$$

Think of a general list $\{b_1, \ldots, b_n\}$ listing the contents of a box. The result $(X_1, \ldots, X_n)$ of exhaustive sampling without replacement is a random permutation of the values in the list, with all $n!$ possible permutations of the indices equally likely. Write $b(k) = b_k$. Then, $X_i = b(Y_i)$ where $(Y_1, \ldots, Y_n)$ is random permutation of $1, 2, \ldots, n$. So

$$X_i = b(Y_i) \text{ where } Y_1, \ldots, Y_n \text{ are exchangeable}$$

But it is intuitively clear (and a consequence of the change of variable principle), that a function b applied to all variables in an exchangeable sequence yields another exchangeable sequence. □

Proof for $2 \le n \le N$ and a general list $\{b_1, \ldots, b_N\}$. For a sample of size n without replacement from a list of N values, the exchangeability follows by viewing the sample of size n as the first n variables in an exhaustive sample, which is exchangeable by the previous case, and appealing to the general fact that subsets of exchangeable variables are exchangeable. □

Examples. The symmetry of sampling without replacement appeared already in Section 2.5, in the derivation of the probability of getting g good elements and b bad elements in a sample of size n without replacement from a population of G good and B bad elements. That calculation used the fact that the probability of getting g

good elements and b bad elements in a particular order is the same for all possible orders. Other consequences of the symmetry appear in Example 1.4.7 and Example 3.1.6. Here are two more examples.

Example 1. Dealing cards.

Five cards are dealt from a standard deck of 52 cards.

Problem 1. What is the probability that the fifth card is a king?

Solution. It is confusing in this problem to think about which of the first four cards are kings. Rather, ignore the first four cards. The fifth card is a card drawn at random from the deck, just like the first card. So the probability that the fifth card is a king is the same as the probability that the first card is a king, that is $1/13$.

Problem 2. What is the chance that the third and fifth cards are black?

Solution. Ignore the first, second, and fourth cards. By the symmetry of sampling without replacement, the third and fifth cards are two cards drawn at random without replacement from the deck, just like the first two cards. So the probability that the third and fifth cards are black is the same as the probability that the first and second cards are black, that is $\frac{26}{52} \times \frac{25}{51}$.

Discussion. This kind of intuitive argument is precisely what is justified by the symmetry of sampling without replacement. Particular problems like these can be solved quickly "by symmetry" without using random variable notation. But the theoretical justification is symmetry of the joint distribution involved.

Example 2. Red and black balls.

Suppose 20 balls are drawn at random without replacement from a box containing 50 red balls and 50 black balls.

Problem 1. What is the probability that the 10th ball is red given that the 18th and 19th balls are red?

Solution. Let X_i be the color of the ith ball drawn. Then $(X_1, X_2, \ldots, X_{20})$ represents a random sample of size 20 without replacement from the population of 100 red and black balls. The problem is to calculate

$$P(X_{10} = red \mid X_{18} = red \text{ and } X_{19} = red) = \frac{P(X_{10} = red \text{ and } X_{18} = red \text{ and } X_{19} = red)}{P(X_{18} = red \text{ and } X_{19} = red)}$$

This conditional probability is determined by the joint distribution of X_{10}, X_{18}, and X_{19}, which is the same as the joint distribution of X_3, X_2 and X_1 by the symmetry of sampling without replacement. So the required probability is the same as

$$P(X_3 = red \mid X_2 = red \text{ and } X_1 = red) = \frac{48}{98}$$

since after drawing two red balls on the first two draws there are 48 red balls remaining out of 98 balls total.

Mean and Variance of the Hypergeometric Distribution

Recall from Section 2.5 the distribution of the number of good elements S_n in a sample of size n for a population of size N containing G good elements:

$$P(S_n = g) = \binom{G}{g}\binom{B}{b}\Big/\binom{N}{n} \quad 0 \le g \le n$$

where $b = n - g$, $B = N - G$ represent numbers of bad elements. The mean and standard deviation of S_n are as follows:

$$E(S_n) = np \text{ and } SD(S_n) = \sqrt{\frac{N-n}{N-1}}\sqrt{npq}$$

where $p = G/N$ is the proportion of good elements in the population, $q = B/N$ the proportion of bad elements in the population. Note that the mean is the same as if the sampling were done with replacement, when the distribution of S_n is binomial(n, p). And the standard deviation is just the familiar binomial standard deviation of $\sqrt{npq}$ multiplied by the factor $\sqrt{\frac{N-n}{N-1}}$, called the *finite population correction factor*.

Proof. Write

$$S_n = I_1 + I_2 + \cdots + I_n,$$

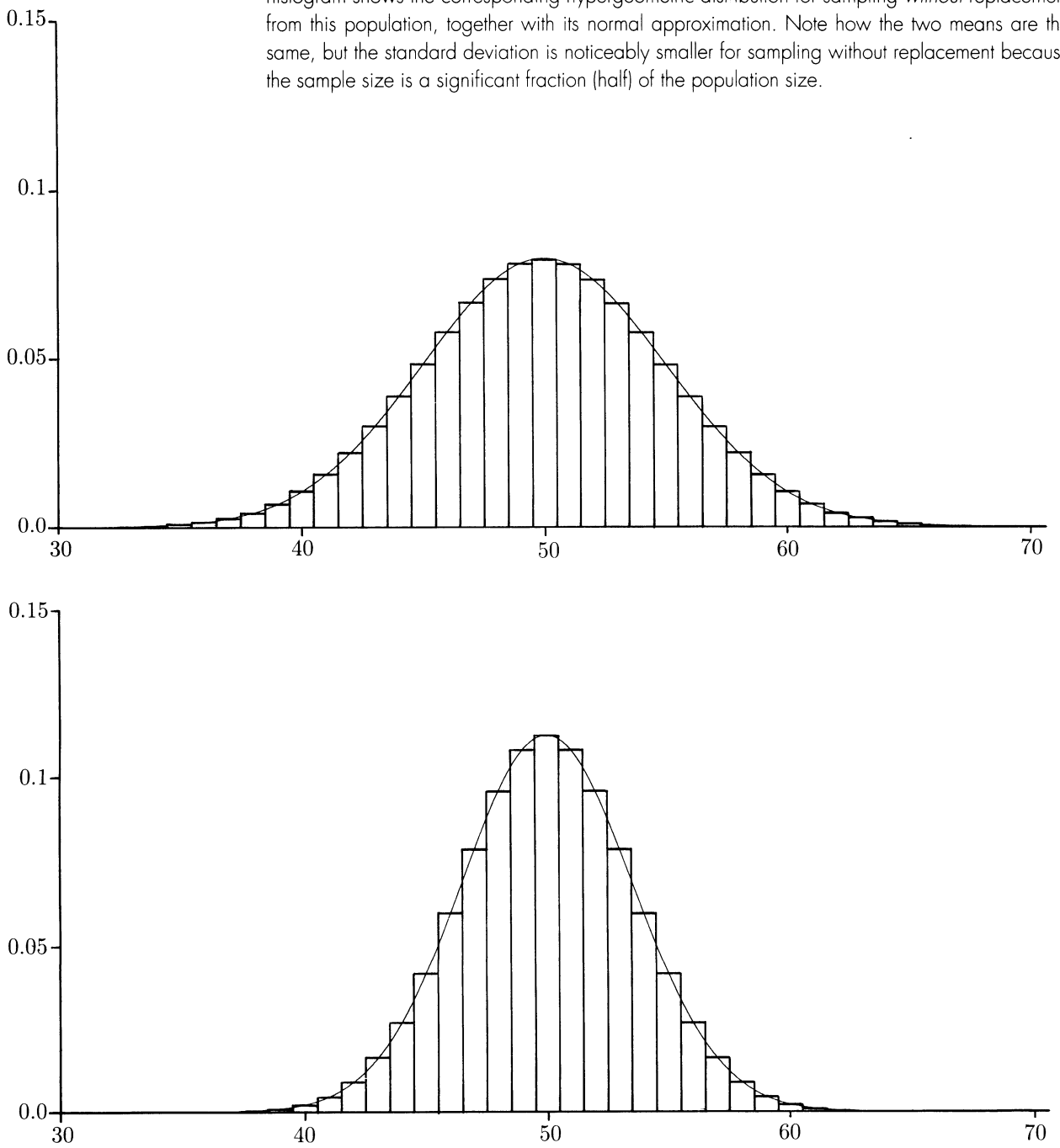
where for each $j = 1, 2, \ldots, n$, I_j is the indicator of the event that the jth draw yields a good element. By the symmetry of sampling without replacement just discussed, the distribution of I_j is the same Bernoulli (G/N) distribution for every j. Thus the expectation of S_n can be computed as

$$E(S_n) = E(I_1) + E(I_2) + \cdots + E(I_n) = nE(I_1) = n\frac{G}{N}$$

The variance can now be computed, starting from a calculation of

$$\begin{aligned} E(S_n^2) &= E[(\sum_j I_j)^2] \\ &= E[\sum_j I_j^2 + 2\sum_{j<k} I_j I_k] \\ &= \sum_j E(I_j^2) + 2\sum_{j<k} E(I_j I_k) \end{aligned}$$

FIGURE 1. **Normal approximation for sampling with and without replacement.** The top histogram shows the binomial $(100, 0.5)$ distribution of the number of good elements in a sample of size $n = 100$ with replacement from a population of size $N = 200$ containing $G = 100$ good elements and $B = 100$ bad ones. The approximating normal curve is superimposed. The bottom histogram shows the corresponding hypergeometric distribution for sampling *without* replacement from this population, together with its normal approximation. Note how the two means are the same, but the standard deviation is noticeably smaller for sampling without replacement because the sample size is a significant fraction (half) of the population size.



$$= n\frac{G}{N} + 2\binom{n}{2}\frac{G}{N}\frac{(G-1)}{(N-1)}$$

because in the first sum there are n identical terms of

$$E(I_j^2) = E(I_j) = \frac{G}{N}$$

(I_j is an indicator variable with value 0 or 1, so $I_j^2 = I_j$) and in the second sum there are $\binom{n}{2}$ identical terms with value

$$E(I_j I_k) = E(I_1 I_2) = \frac{G}{N} \cdot \frac{(G-1)}{(N-1)}$$

the probability of getting good elements on two consecutive draws, since $I_1 I_2$ is one if both I_1 and I_2 are 1, and 0 otherwise. Now use

$$Var(S_n) = E(S_n^2) - [E(S_n)]^2$$

and simplify to obtain the expression for $SD(S_n) = \sqrt{Var(S_n)}$. □

Remark. A similar argument shows that the same finite population correction factor applies for sums or averages of other kinds of variables in sampling without replacement, not just indicator variables. See Example 6.4.7.

The normal approximation. This can be used for sampling without replacement exactly as in the binomial case for sampling with replacement, provided the finite population correction factor is used for the standard deviation. The approximation is good provided the standard deviation is sufficiently large. This can be shown by consideration of consecutive odds ratios, just as in the binomial case. See Figure 1 for an illustration.

Exercises 3.6

1. Five cards are dealt from a standard deck of 52. Find
   a) the probability that the third card is an ace;
   b) the probability that the third card is an ace given the last two cards are not aces;
   c) the probability that all cards are of the same suit;
   d) the probability of two or more aces.

2. **Cards.** A deck of 52 cards is shuffled and dealt. Find the probabilities of the following events:
   a) the tenth card is a queen;
   b) the twentieth card is a spade;
   c) the last five cards are spades;

d) The last king appears on the 48th card.

3. Conditional probabilities. In the setting of Exercise 2, denote by A, B, C, and D the events defined in parts a), b), c) and d) of that exercise. Find:

a) $P(B|C)$; b) $P(C|B)$; c) $P(B|A)$; d) $P(A|B)$; e) $P(D|C)$; f) $P(C|D)$;

4. Testing for defectives. Suppose a lot of 5 items contains two defective items. The items are tested one by one in random order. Let T_1 be the number of the test on which the first defective item is discovered, and T_2 the number of the test on which the second is discovered.

a) Display the distribution table of T_1.

b) Without further calculation, display the distribution table of $6 - T_2$.

c) Without further calculation, display the distribution table of T_2.

d) Display the joint distribution table of T_1 and T_2.

e) Are the random variables T_1, $T_2 - T_1$, $6 - T_2$ exchangeable? Prove your answer.

f) Find the distribution of $T_2 - T_1$.

5. Suppose n balls are thrown independently at random into b boxes. Let X be the number of boxes left empty. Use the method of indicators to find expressions for $E(X)$ and $Var(X)$.

6. Mean and SD of the number of matches. There are n balls labeled 1 through n, and n boxes labeled 1 through n. The balls are distributed randomly into the boxes, one in each box, so that all $n!$ permutations are equally likely. Say that a match occurs at place i if the ball labeled i happens to fall in the box labeled i. Let M be the total number of matches.

a) Find $E(M)$. b) Find $SD(M)$.

c) For very large n, what do you think is the approximate distribution of M? Give an intuitive explanation for your answer. Check that your answer makes sense in view of your answers to a) and b) and the answer to Exercise 28 from the Chapter 2 Review Exercises.

7. Suppose n cards are dealt from a standard deck of 52 cards. Calculate a) the expectation and b) the variance of the number of red cards among the n cards dealt.

8. A deck of 52 cards is shuffled and split into two halves. Let X be the number of red cards in the first half. Find: a) a formula for $P(X = k)$;

b) $E(X)$; c) $SD(X)$; d) $P(X \geq 15)$, approximately, using the normal curve.

9. A population contains G good and B bad elements, $G + B = N$. Elements are drawn one by one at random without replacement. Suppose the first good element appears on draw number X. Find simple formulae, not involving any summation from 1 to N, for:

a) $E(X)$; b) $SD(X)$.

[*Hint:* Write $X - 1$ as a sum of B indicators.]

10. Success runs in sampling without replacement. Repeat Exercise 3.2.22 for the random sequence of successes and failures obtained by a sampling n times without replacement from a population of G good and $N - G$ bad elements, where each draw of a good element is called success, and each draw of a bad element a failure.

11. Sampling without replacement. Let X_j be the indicator of the event that a good element appears at place j in a random ordering of n elements consisting of g good elements and $n - g$ bad ones.

a) Find a formula for $P(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$.

b) Are the random variables $X_1, \ldots, X_n$ independent? Prove your answer.

c) Are they exchangeable? Prove your answer.

12. Discrete order statistics. In an exhaustive random sample without replacement of a population of N elements, containing n good and $N - n$ bad elements, let $1 \leq T_1 < T_2 < \cdots < T_n \leq N$ denote when the good elements appear. Part d) of this exercise explains why the random variables $T_1, \ldots, T_n$ with possible values in $\{1, \ldots, N\}$ are discrete analogs of the order statistics of n independent uniform $(0, 1)$ variables, studied in Section 4.6.

a) Show that $\{T_1, \ldots, T_n\}$, the random set of times when good elements appear, is uniformly distributed over all subsets of n elements of $\{1, \ldots, N\}$. That is to say, the set of times when good elements appear is a simple unordered random sample of size n from $\{1, \ldots, N\}$.

b) Find a formula for $P(T_1 = t_1, \ldots, T_n = t_n)$ for $1 \leq t_1 < t_2 < \cdots < t_n \leq N$.

c) Use a counting argument to find a formula for $P(T_i = t)$ for each $i = 1, \ldots, n$ and $t = 1, \ldots, N$.

d) Let $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$ denote the order statistics, that is, the values in increasing order, of n independent trials $U_1, \ldots, U_n$ with uniform distribution on $\{1, \ldots, N\}$. Let D denote the event that the $U_i, 1 \leq i \leq n$ are all distinct. Show that the conditional joint distribution of $U_{(1)}, \ldots, U_{(n)}$ given D is identical to the joint distribution of $T_1, \ldots, T_n$ found in part b). What is $P(D)$? Show that $P(D) \to 1$ as $N \to \infty$ for fixed n.

[It follows that for fixed n, as $N \to \infty$, the limiting joint distribution of $(T_1, \ldots, T_n)/N$ is the joint distribution of the order statistics of n independent uniform $(0, 1)$ random variables. In particular, part c) implies the asymptotic distribution of T_i/N is the beta $(i, n - i + 1)$ distribution, as obtained directly from the continuous model in Section 4.6. A number of interesting results for continuous uniform order statistics can be derived via this passage to the limit. See Chapter 6 Review Exercises 31, 32, and 33.

13. Discrete spacings. As in Exercise 12, let $T_1 < \ldots < T_n$ be the places that good elements appear in a random ordering of n good and $N - n$ bad elements. (In terms of a shuffled deck of N cards with n aces, T_i represents the place in the deck where the ith ace lies.) Let $W_1 = T_1 - 1$, the number of bad elements before the first good one. For $2 \leq i \leq n$, let $W_i = T_i - T_{i-1} - 1$, the number of bad elements between the $(i - 1)$th and ith good ones. Let $W_{n+1} = N - T_n$, the number of bad elements after the last good one. Think of the W_i as spacings between the good elements.

a) Find the joint distribution of $W_1, \ldots, W_{n+1}$.

b) Show that the $n+1$ random variables $W_1, \ldots, W_{n+1}$ are exchangeable, hence identically distributed, but not independent.

c) Find a formula for $P(W_i = w)$ for $0 \le w \le N$.

d) Find $E(W_i)$ for $1 \le i \le n+1$ and $E(T_i)$ for $1 \le i \le n$. [*Hint*: Use the symmetry.] Evaluate in the case $N = 52$ and $n = 4$ to find the mean number of cards between any two aces, and the mean position in the deck of the ith ace. (See Chapter 6 Review Exercise 29 for the variance.)

e) Show that for $1 \le i < j \le n+1$ the random variable $W_i + W_j$ has the same distribution as $T_2 - 2$. Deduce from Exercise 12c) a formula for $P(W_i + W_j = t)$ for $0 \le t \le N$.

f) Let $D_n = T_n - T_1 - 1$, the number of elements between the first and last good elements (including the other $n-2$ good ones). Use the result of e) to find a formula for $P(D_n = d)$, $0 \le d \le N$, and find $E(D_n)$.

14. Consecutive pairs. Consider a well-shuffled deck of N cards, with n aces and $N-n$ non-aces.

a) Show by a counting argument that the probability that there are at least two consecutive aces somewhere in the deck is $1 - \binom{N-n+1}{n} \Big/ \binom{N}{n}$ [*Hint*: Look for a one-to-one correspondence].

b) Check the above formula by more direct counting arguments in each of the following three special cases: $n = 2, N = 2n-1$, and $N = 2n$.

For the following parts, assume a standard deck of 52 cards, and evaluate the probabilities of the events as decimals:

c) The ace of spades is next to the ace of clubs.

d) There are at least two consecutive aces somewhere in the deck.

e) There are at least two consecutive spades somewhere in the deck.

f) There is no pair of adjacent black cards anywhere in the deck.

15. Runs and Spacings. As in Exercise 13 let $W_1, W_2, \ldots, W_{n+1}$ be the exchangeable sequence of spacings defined by a random ordering of n aces and $N-n$ non-aces.

a) Explain why the probability evaluated in Exercise 14, that there are at least two consecutive aces somewhere in the deck, is

$$1 - P(W_i \ge 1 \text{ for every } 2 \le i \le n)$$

b) Show that for any sequence of $n+1$ non-negative integers $t_1, \ldots, t_{n+1}$ with $t_1 + \cdots + t_{n+1} = t$,

$$P(W_i \ge t_i \text{ for every } 1 \le i \le n+1) = \binom{N-t}{n} \Big/ \binom{N}{n}$$

c) What special case of b) yields the result of Exercise 14?

16. Distribution of the longest run. As in Exercises 13 and 15, let $W_1, W_2, \ldots, W_{n+1}$ be the exchangeable sequence of spacings defined by a random ordering of n aces and $N-n$ non-aces. Let $W_{\max} = \max_i W_i$ where the max is over $1 \le i \le n+1$. So $W_{\max}$ is the length of the longest run of non-aces in the deck.

a) Show by using the result of Exercise 15, and the inclusion–exclusion formula of Exercise 1.3.12 that

$$P(W_{\max} \geq r) = \sum_{i=1}^{n+1} (-1)^{i-1} \binom{n+1}{i} \binom{N-ir}{n} \Big/ \binom{N}{n}$$

b) Denote the above expression for $P(W_{\max} \geq r)$, which depends on N, n, and r, by $P(N, n, r)$. Let S_N be the number of successes in N Bernoulli (p) trials and R_N be the longest run of successes in the N trials. Explain why

$$P(R_N \geq r | S_N = k) = P(N, N-k, r)$$

and why this conditional probability does not depend on p.

c) Show that the probability that there is a run of at least r consecutive successes in N Bernoulli (p) trials is

$$P(R_N \geq r) = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} P(N, N-k, r)$$

d) Find as a decimal the probability that the longest run of heads in 10 fair coin tosses is exactly r for each $0 \leq r \leq 10$. What is the most likely length of the longest run? What is the expected length of the longest run?

e) What is the probability that there is a run of either at least 5 heads or at least 5 tails in 10 fair coin tosses?

Random Variables: Summary

Random variable X: symbol representing an outcome.

Range of X: set of all possible values of X.

Distribution of X: The probability distribution over the range of X defined by probabilities $P(X = x)$ for x in the range of X.
$P(X \in B) = \sum_{x \in B} P(X = x)$ for B a subset of the range of X.

Change of variable formula: $P(f(X) = y) = \sum_{x:f(x)=y} P(X = x)$ gives the distribution of a function $f(X)$ in terms of the distribution of X.

Joint outcome (X, Y): $P(x, y) = P(X = x, Y = y)$

$$P(X = x) = \sum_{\text{all } y} P(x, y) \qquad P(X < Y) = \sum_{x} \sum_{y>x} P(x, y)$$

Equality of random variables: $X = Y$ means $P(X = Y) = 1$.

Equality in distribution: X and Y have the same distribution if $P(X = x) = P(Y = x)$ for all x in the range of X (= range of Y). If $X = Y$ then X and Y have the same distribution, but not conversely.

Independence: For n random variables

$$P(X_1{=}x_1, X_2{=}x_2, \ldots, X_n{=}x_n) = P(X_1{=}x_1)P(X_2{=}x_2) \cdots P(X_n{=}x_n)$$

for all possible values x_i of X_i, $i = 1, \ldots, n$,
- functions of disjoint blocks of independent random variables are independent.

Expectation: $E(X) = \sum_x xP(X = x)$

- average value of X weighted by probabilities;
- long-run average value of independent variables with same distribution as X;
- center of mass of distribution of X
- properties: generalize properties of averages: see summary on pages 180 – 181

Variance: $Var(X) = E(X-\mu)^2 = E(X^2) - \mu^2$ where $\mu = E(X)$.

Standard deviation: $SD(X) = \sqrt{Var(X)}$: measure of spread in the distribution of X.

Scaling: $Var(aX+b) = a^2\,Var(X)$, $SD(aX+b) = |a|SD(X)$.

Chebychev's inequality: $P\left[|X - E(X)| > kSD(X)\right] \le \dfrac{1}{k^2}$

Sums: For independent random variables $X_1, \ldots, X_n$, if $S_n = X_1 + \cdots + X_n$,

$$\begin{aligned} Var(S_n) &= Var(X_1) + \cdots + Var(X_n) \\ &= nVar(X_1) \quad \text{if the } X_i \text{ all have same distribution.} \\ \text{Compare} \quad E(S_n) &= E(X_1) + \cdots + E(X_n) \quad \text{(true even if dependent)} \\ &= nE(X_1) \quad \text{if the } X_i \text{ all have same distribution.} \end{aligned}$$

Square root law: For independent X_i with same distribution, S_n as above, and $\bar{X}_n = S_n/n$ the average

$$SD(S_n) = SD(X_1)\sqrt{n} \qquad SD(\bar{X}_n) = SD(X_1)/\sqrt{n}$$

Law of averages: $\bar{X}_n$ is nearly certain to be close to $E(X_1)$ for large n.

Normal approximation: For S_n as above, with $E(X_i) = \mu, \mathrm{SD}(X_i) = \sigma$,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{(\bar{X}_n - \mu)\sqrt{n}}{\sigma}$$

has distribution which approaches standard normal as $n \to \infty$, no matter what the common distribution of the X_i.

Infinite sum rule. If event A splits into an infinite sequence of mutually exclusive cases $A_1, A_2, A_3, \ldots$, so $A = A_1 \cup A_2 \cup A_3 \cup \cdots$, where $A_i \cap A_j = \emptyset$, $i \ne j$, then

$$P(A) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

Discrete distribution on {0, 1, 2, ...}: defined by a sequence of probabilities $p_0, p_1, p_2, \ldots$ such that $p_i \ge 0$ for all i, and $\sum_i p_i = 1$.

Geometric, negative binomial, and Poisson distributions.
See Distribution Summaries on pages 476 – 488.

Review Exercises

1. A fair die is rolled ten times. Write down numerical expressions for:

a) the probability of at least one six in the ten rolls;

b) the expected number of sixes in the ten rolls;

c) the expected sum of the numbers in the ten rolls;

d) the probability of 2 sixes in the first five rolls given 4 sixes in the ten rolls;

e) the probability of getting strictly more sixes in the second five rolls than in the first five.

2. A fair die is rolled repeatedly. Calculate, correct to at least two decimal places:

a) the chance that the first 6 appears before the tenth roll;

b) the chance that the third 6 appears on the tenth roll;

c) the chance of seeing three 6's among the first ten rolls, given that there were six 6's among the first twenty rolls;

d) the expected number of rolls until six 6's appear;

e) the expected number of rolls until all six faces appear.

3. Two fair dice are rolled independently. Let X be the maximum of the two rolls, and Y the minimum.

a) What is $P(X = x)$ for $x = 1, \ldots, 6$?

b) What is $P(Y = y|X = 3)$ for $y = 1, \ldots, 6$?

c) What is the joint distribution of X and Y?

d) What is $E(X + Y)$?

4. Let X and Y be independent, each uniform on $\{0, 1, \ldots, 100\}$. Let $S = X + Y$. For $n = 0, \ldots, 200$, find:

a) $P(S = n)$; b) $P(S \leq n)$. c) Sketch graphs of these functions of n.

5. Someone plays roulette the following way: before each spin he rolls a die, and then he bets on red as many dollars as there were spots on the die. For example, if there were 4 spots he bets \$4.
If red comes up he gets the stake back plus an amount equal to the stake. If red does not come up he loses the stake. In the example above, if red comes up he gets the stake of \$4 back plus an additional \$4. If red does not come up he loses his stake of \$4. The probability of red coming up is 18/38.

a) What is his expected gain on one spin?

b) What is the expected number of spins it will take until red comes up for the first time?

c) What is the expected number of spins it will take until the first time the person bets exactly \$4 on one spin and wins.

6. A gambler repeatedly bets 10 dollars on red at a roulette table, winning 10 dollars with probability 18/38, losing 10 dollars with probability 20/38. He starts with capital of 100 dollars, and can borrow money if necessary to keep in the game.

a) Find exact expressions for the probabilities that after 50 plays the gambler is:

i) ahead; ii) not in debt.

b) Find the mean and variance of the gambler's capital after 50 plays.

c) Use the normal approximation to estimate the probabilities in a) above.

7. Suppose an airline accepted 12 reservations for a commuter plane with 10 seats. They know that 7 reservations went to regular commuters who will show up for sure. The other 5 passengers will show up with a 50% chance, independently of each other.

a) Find the probability that the flight will be overbooked, i.e., more passengers will show up than seats are available.

b) Find the probability that there will be empty seats.

c) Let X be the number of passengers turned away. Find $E(X)$.

8. A box contains w white balls and b black balls. Balls are drawn one by one at random from the box, until b black balls have been drawn. Let X be the number of draws made. Find the distribution of X,

a) if the draws are made with replacement;

b) if the draws are made without replacement.

9. **The doubling cube.** A doubling cube is a die with faces marked $2, 4, 8, 16, 32$, and 64. Suppose two doubling cubes are rolled. Let XY be the product of the two numbers. Find a) $P(XY < 100)$; b) $P(XY < 200)$; c) $E(XY)$; d) $SD(XY)$.

10. **Matching.** Suppose each of n balls labeled 1 to n is placed in one of n boxes labeled 1 to n. Assume the n placements are made independently and uniformly at random (so each box can contain more than one ball). A match occurs at place k if ball number k falls in box k. Find:

a) the probability of a match at i and no match at j;

b) the expected number of matches.

11. Data for performances of a particular surgical operation show that two operations per thousand have resulted in the death of the patient. Let X be the number of deaths due to the next thousand operations of this kind. Which of these three numbers is the smallest and which is the largest

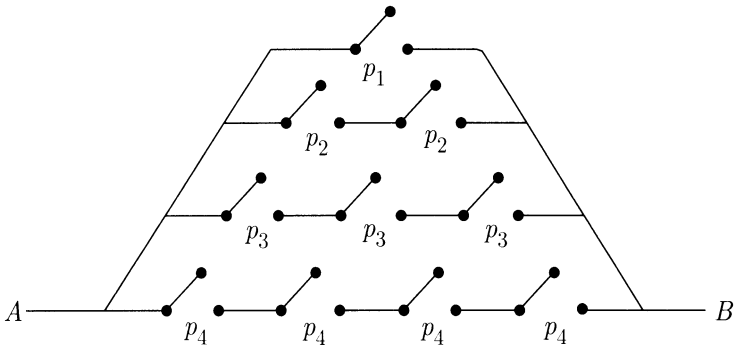
$$P(X < 2), \quad P(X = 2), \quad P(X > 2)?$$

Explain carefully the assumptions of your answer.

12. Consider an unlimited sequence of independent trials resulting in success with probability p, failure with probability q. For $s = 1, 2, \ldots, f = 1, 2, \ldots$ calculate the probability that s successes in a row occur before f failures in a row. [*Hint:* Let A be the event in question, $P_1 = P(A \mid \text{first trial a success})$, and $P_0 = P(A \mid \text{first trial a failure})$. Given the first trial is a success, for A to occur, either the next $s - 1$ trials must be successes, or the first failure must come at the tth trial for some $2 \le t \le s$, then subsequently the event A must occur starting from a failure. This gives one equation relating P_1 to P_0. Find another by conditioning on the first trial being a failure, then solve for P_0 and P_1, hence $P(A)$.]

13. Let X and Y be independent random variables with $E(X) = E(Y) = \mu$, $Var(X) = Var(Y) = \sigma^2$. Show that $Var(XY) = \sigma^2(2\mu^2 + \sigma^2)$.

14. A circuit contains 10 switches, arranged as in the figure below. Assume switches perform independently of each other, and are closed with probabilities indicated in the figure. Current flows through a switch if and only if it is closed.



a) What is the probability that current flows between points A and B?

b) Find the mean and standard deviation of the number of closed switches.

15. A roulette wheel is spun independently many times. On each spin the chance of a seven appearing is $1/38$.

a) What is the exact distribution of the number of sevens in the first 100 spins?

b) Give a simple approximation for this distribution.

c) What is the distribution of the number Z of spins required to produce three sevens?

d) What is $E(Z)$?

16. Random products mod 10. Pick two successive digits from a table of random digits from $\{0, 1, \ldots, 9\}$. Multiply them together, and let D be the last digit of this random product. For example,

$$(3, 9) \to 27 \to 7$$

$$(2, 4) \to 8 \to 8$$

Find the distribution of D, and calculate its mean.

17. Suppose N dice are rolled, where $1 \leq N \leq 6$.

a) Given that no two of the N dice show the same face, what is the probability that one of the dice shows a six? Give a formula in terms of N.

b) In a) the number of dice N was fixed, but now repeat assuming instead that N is random, determined as the value of another die roll. Your answer now should be simply a number, not involving N.

18. Expected number of records. Suppose 100 cards numbered 1 to 100 are shuffled and dealt one by one.

a) What is the fair price to pay in advance if you receive one cent for the first card and then one cent for each card dealt whose number is greater than those of all previous cards dealt?

b) If you paid 10 cents for each play of this game, and played 25 times (meaning you paid a total of 250 cents for 25 separate deals of the 100 card deck) what, approximately, is the chance that you would come out ahead?

19. Suppose that X has Poisson (μ) distribution, and that Y has geometric (p) distribution on $\{0, 1, 2, \dots\}$ independently of X.

a) Find a formula for $P(Y \geq X)$ in terms of p and μ.

b) Evaluate numerically for $p = 1/2$ and $\mu = 1$.

20. a) Show that for all p between 0 and 1: $p(1-p) \leq 1/4$.

b) A certain university has about 12,000 students. To estimate the percentage of students who have part-time jobs, someone takes a random sample from a list of all students in the university. How big does the sample need to be so that the margin of error in the estimate (i.e., the standard deviation of the percentage in the sample) is at most 5%?

21. Suppose X and Y are independent with $P(X = j) = p(1-p)^j$ for $j = 0, 1, \dots$ and $P(Y = k) = (k+1)p^2(1-p)^k$ for $k = 0, 1, \dots$. Find the distribution of $Z = X + Y$. [*Hint*: Represent X and Y in terms of a biased coin-tossing sequence.]

22. **The newsboy problem.** A newsboy buys papers at 10 cents a copy and sells them on the street corner at 25 cents a copy. He must buy all his papers at once, but he can sell only as many as are demanded on the street. Left-over papers are a dead loss. Over the last few years, demand has been fluctuating at around 100 papers per day. He has been buying 100 papers and selling them all about half the time. Assuming that the demand for papers has an approximately Poisson distribution, find:

a) the newsboy's long-run average profit per day;

b) how many papers the newsboy should buy each day to maximize his long-run average profit.

23. Suppose you economize your use of toothpicks by breaking whole toothpicks in half and only using half at a time. Starting from a full box of n toothpicks, you draw repeatedly at random from the box. In case you draw a whole toothpick, you use half and throw it away, and replace the other half. In case you draw half a toothpick, you use it and throw it away. So the box will be empty after exactly $2n$ draws. Suppose that on any draw, each whole toothpick in the box has the same chance of being drawn, and so does each half toothpick, but the halves have half the chance of the wholes. Let H be the random number of half toothpicks remaining in the box after the last whole toothpick has been drawn and half of it replaced. So H has possible values between 1 (e.g., if you draw alternately whole–half–whole–half ...) and n (e.g., if you draw n wholes in a row, followed by n halves).

a) Find a formula for $P(H = k), \quad k = 1, 2, \dots, n$.

b) What happens to the distribution of H as $n \to \infty$?

c) Find an asymptotic formula for $E(H)$ as $n \to \infty$.

d) If you start with $n = 100$ toothpicks, about how many halves do you expect to be left with?

e) For $n = 100$, find a and b so that $P(a \leq H \leq b) \approx 95\%$ with $b - a$ as small as possible.

24. The voter paradox.

a) Can random variables X, Y, and Z be such that each of the three probabilities $P(X > Y)$, $P(Y > Z)$, and $P(Z > X)$, is strictly greater than $\frac{1}{2}$? [*Hint*: Try a joint distribution of X, Y, and Z which is uniform on some of the 6 permutations of $(1, 2, 3)$.]

b) What is the largest that the minimum of the above three probabilities can possibly be? Prove your answer. [*Hint*: The sum of the probabilities is an expectation.]

c) A survey is conducted to determine the popularity of three candidates A, B, and C. Each voter is asked to rank the candidates in order of preference. When the results are analyzed, it is found that more than 50% of the voters prefer A to B, more than 50% prefer B to C, and more than 50% prefer C to A. How is this possible? Explain carefully the connection to previous parts.

d) Generalize a) and b) to $n \geq 3$ random variables instead of $n = 3$.

e) Repeat a) for independent X, Y, and Z. [*Hint*: Try $P(X = 5) = p_1$, $P(X = 2) = 1 - p_1$, $P(Y = 4) = p_2$, $P(Y = 1) = 1 - p_2$, and $P(Z = 3) = 1$. Deduce that the three probabilities can all be as large as the golden mean $(-1 + \sqrt{5})/2$. This is known to be the largest possible for independent variables, but I don't know the proof.]

25. Let Y_1 and Y_2 be independent random variables each with probability distribution defined by the following table:

value	0	1	2
probability	1/2	1/3	1/6

a) Display the probability distribution of $Y_1 + Y_2$ in a table. Express all probabilities as multiples of $1/36$.

b) Calculate $E(3Y_1 + 2Y_2)$.

c) Let X_1 and X_2 be the numbers on two rolls of a fair die. Define a function f so that $(f(X_1), f(X_2))$ has the same distribution as (Y_1, Y_2).

26. The horn on an auto operates on demand 99% of the time. Assume that each time you hit the horn, it works or fails independently of all other times.

a) How many times would you expect to be able to honk the horn with a 50% probability of not having any failures?

b) What is the expected number of times you hit the horn before the fourth failure?

27. A certain test is going to be repeated until done satisfactorily. Assume that repetitions of the test are independent and that each has probability 0.25 of being satisfactory. The first 5 tests cost \$100 each to perform and thereafter cost \$40 each, regardless of the outcomes. Find the expected cost of running the tests until a satisfactory result is obtained.

28. Let $X_1, X_2, \ldots$ be a sequence of independent trials, and suppose that each X_i has distribution P_1 over some range space Ω_1. Let $W_1, W_2, \ldots$ be the successive waiting times between trials s such that X_s is in A, where A is some subset Ω_1, and let $Y_1, Y_2, \ldots$ be the successive values in A which appear at trials W_1, $W_1 + W_2$, $W_1 + W_2 + W_3, \ldots$.

a) Show that $W_1, W_2, \ldots$, $Y_1, Y_2, \ldots$ are independent random variables, the W's all having geometric distribution on $\{1, 2, \ldots\}$ with parameter $P_1(A)$, and the Y's all having the distribution P_1 conditioned on A.

b) Deduce from the law of large numbers the long run frequency interpretation of $P_1(B|A)$ as the limiting proportion of those trials which are A's that turn out also to be B's.

29. Polya's urn scheme. (Continuation of Exercise 1.5.2). An urn contains w white and b black balls. A ball is drawn from the urn, then replaced along with d more balls of the same color. So after n such draws with multiple replacement, the urn contains $w+b+nd$ balls. Let $X_i = 1$ if the ith ball drawn is black and $X_i = 0$ if the ith ball drawn is white.

a) Find a formula for the probability $P(X_1 = x_1, \ldots, X_n = x_n)$ in terms of w, b, d, n and k, where $k = x_1 + \cdots + x_n$ is the number of 1's in the sequence $(x_1, \ldots, x_n)$.

b) Let $S_n = X_1 + \cdots + X_n$. What does S_n represent? Find a formula for $P(S_n = k)$ for $0 \leq k \leq n$.

c) What is the distribution of S_n in the special case $b = w = d = 1$?

d) Are $X_1, \ldots, X_n$ independent? Are they exchangeable? (Refer to Section 3.6.)

e) Find a formula for $P(X_n = 1)$, the probability of a black ball on draw n, in terms of b,w,d, and n. [*Hint*: The probability does not depend on all of the parameters.]

f) Find the probability that the fifth ball drawn is black given that the tenth ball drawn is black.

30. Diagonal neighbor random walk. Let (S_n, T_n) denote the position after n steps of a random walk on the lattice of points in the plane with integer coordinates, starting from $(S_0, T_0) = (0, 0)$. Suppose that $S_{n+1} = S_n \pm 1$ and $T_{n+1} = T_n \pm 1$ where the signs are picked by two independent tosses of a fair coin, independently at each step.

a) For $c > 0$, find the limit as $n \to \infty$ of the probability that (S_n, T_n) is inside the square with corners at $(\pm c\sqrt{n}, \pm c\sqrt{n})$.

b) Let $R_n = \sqrt{S_n^2 + T_n^2}$, the distance from the origin. Find $E(R_n^2)$.

c) Find b, as small as you can, such that $E(R_n) \leq \sqrt{bn}$ for every n.

d) Let p_n denote the probability that the random walk is at $(0, 0)$ after n steps. Find p_4 as a decimal.

e) Show that $p_{2m} \sim c/m$ as $m \to \infty$ for a constant c. What is c?

31. Nearest neighbor random walk. Let (S_n, T_n) be the position after n steps of a random walk as in the previous exercise, but now instead of diagonal moves, suppose at each step the move is made with equal probability up, down, left or right, to one of the four nearest neighbors in the lattice. For $c > 0$, find the limit as $n \to \infty$ of the probability that $|S_n| < c\sqrt{n}$. The events $|S_n| < c\sqrt{n}$ and $|T_n| < c\sqrt{n}$ are clearly not independent for this random walk, but they turn out to be approximately independent for large n. Assuming the error of this approximation tends to zero as $n \to \infty$ (something not

obvious, but true: see Chapter 5 Review Exercise 31 for an explanation), repeat part a) of the previous exercise for this random walk. Now repeat the rest of the previous exercise for this random walk.

32. King's random walk. Same as Exercise 30, but now make each move like a king on an infinite chessboard, with equal probabilites to the 8 nearest or diagonal neighbors. [The two components are still asymptotically independent. This can be proved for any step distribution with mean zero and uncorrelated components, that is to say $E(S_1T_1) = 0$.]

33. From a very large collection of red and black balls, half of them red and half black, I pick n balls at random and put these n balls in a bag. Suppose you now draw k balls from the bag, with replacement and mixing of the balls between draws.

a) Show that given that all k balls you pick are red, the chance that the n balls in the bag are all red is

$$P(n \text{ red in bag} \,|\, \text{pick } k \text{ red}) = \frac{n^k}{2^n E(X^k)}$$

where X is a binomial $(n, 1/2)$ random variable.

b) Simplify this expression further in the cases $k = 1$ and $k = 2$.

c) Find a similar formula assuming instead that the sample of size k is drawn from the bag *without* replacement. Deduce by calculating the same quantity in a different way that

$$E(X)_k = (n)_k/2^k,$$

where $(X)_k = X(X-1)\cdots(X-k+1)$.

d) Use the identity of c) to simplify the answer to a) in case $k = 3$.

e) Show by a variation of the above calculations that for a binomial (n, p) random variable X,

$$E(X)_k = (n)_k\, p^k.$$

Check that for $k = 1$ and 2 this agrees with the formulae for $E(X)$ and $Var(X)$.

34. Probability generating functions. For a random variable X with non-negative integer values, let $G_X(z) = \sum_{i=0}^{\infty} P(X = i)z^i$, be the probability generating function of X, defined for $|z| < 1$. (Refer to Exercises 3.4.22, 3.4.23 and 3.5.19.) Show that:

a) $G_X(z) = E(z^X)$.

b) If X and Y are independent, then $G_{X+Y}(z) = G_X(z)G_Y(z)$. That is to say, $P(X + Y = k)$ is the coefficient of z^k in $G_X(z)G_Y(z)$.

Generalize the above result to obtain the probability generating function of $S_n = X_1 + \cdots + X_n$ for independent X_i. Now identify the generating function and hence the distribution of S_n in case the distribution of the X_i is c) binomial (n_i, p);

d) Poisson (μ_i); e) geometric (p); f) negative binomial (r_i, p);

35. Binomial moments and the inclusion–exclusion formula. Let X be the number of events that occur in some collection of events $A_1, \ldots, A_n$. So $X = \sum_j I_j$ where I_j is the indicator of A_j.

a) Explain the identity of random variables $\binom{X}{2} = \sum_{i<j} I_i I_j$. [*Hint*: Think in terms of a gambler who for every $i < j$ bets that both A_i and A_j will occur. If the number of events that occurs is, say x, how many bets has the gambler won?]

b) For $k = 0, 1, \ldots, n$ the kth *binomial moment* of X is $b_k = E[\binom{X}{k}]$. Show:

$$b_1 = \sum_i P(A_i); \qquad b_2 = \sum_{i<j} P(A_i A_j); \qquad b_3 = \sum_{i<j<k} P(A_i A_j A_k) \quad \text{and so on.}$$

c) Notice that these are the sums of probabilities that appear in the inclusion–exclusion formula from Exercise 1.3.12. Note also that $b_0 = 1$. Deduce that

$$P(X = 0) = \sum_{k=0}^{n} (-1)^k b_k$$

d) **Sieve formula.** [*Hard.*] Show that for every $m = 1, 2, \ldots n$

$$P(X = m) = \sum_{k=m}^{n} \binom{k}{m} (-1)^{m-k} b_k \text{ and } P(X \geq m) = \sum_{k=m}^{n} (-1)^{k-m} \binom{k-1}{m-1} b_k$$

[*Hint*: $P(X = m)$ is the coefficient of z^m in the probability generating function $G_X(z)$ (see Exercise 3.4.22). Consider the Taylor series of $G_X(z)$ about 1, and use the fact that $G_X(z)$ is a polynomial.]

36. Moments of the binomial distribution. Let S_n be the number of successes in n Bernoulli (p) trials.

a) Use the formula for binomial moments in Exercise 35 to find a simple formula for the kth binomial moment of S_n.

b) Check that your formula implies the usual formulae for the mean and variance, and the formula of Exercise 3.3.33 for the skewness of the binomial (n, p) distribution of S_n.

37. Binomial moments of the hypergeometric distribution. Let S_n be the number of good elements in a sample of size n without replacement from a population of G good and $N - G$ bad elements.

a) Use the formula for binomial moments in Exercise 35 to find a formula for the kth binomial moment of S_n for $k = 1, 2, 3$.

b) Check that your formula implies the formulae of this section for the mean and variance.

c) Find the skewness of the distribution of S_n.

38. Limit distribution for the number of matches. Let M_n denote the number of matches in the matching problem of Chapter 2 Review Exercise 28, for a random permutation of n items.

a) Use the method of Exercise 35 to find the kth factorial moment of M_n.

b) Show that for $1 \leq k \leq n$ this kth factorial moment is identical to the kth factorial moment of the Poisson (1) distribution.

c) Show that for $1 \leq k \leq n$ the ordinary kth moment of M_n equals the ordinary kth moment of the Poisson (1) distribution. Deduce that for every k, as $n \to \infty$, the kth moment of the distribution of M_n converges to the kth moment of the Poisson (1) distribution.

d) It is known (though not easy to prove) that if all the moments of a sequence of distributions P_n on $\{0, 1, \ldots\}$ converge to those of a Poisson (λ) distribution, then for every $k = 1, 2, \ldots$, $P_n(k)$ converges to the Poisson (λ) probability of k. In the present problem, this implies that as $n \to \infty$, the limiting distribution of M_n is Poisson (1): $P(M_n = k) \to e^{-1}/k!$. Deduce this result another way by applying part a) and the sieve formula of Exercise 35.

39. Recovering a distribution over {0,1,...,n} from its moments. Let X be a random variable with possible values $\{0, 1, \ldots, n\}$. Assuming the results of Exercise 35, show

a) For some coefficients $c_{n,k}$ not depending on the distribution of X, (which you need not determine explicitly)

$$P(X = 0) = \sum_{k=0}^{n} c_{n,k} E[X^k]$$

b) Find the values of $c_{n,k}$ for $0 \leq k \leq n \leq 3$.

c) Show that for every $m = 1, \ldots n$, the probability $P(X = m)$ can be expressed as a linear combination (which you need not determine explicitly) of the first n ordinary moments of X. [Exercise 40 gives a generalization.]

40. Recovering a distribution on n values from its moments. For a random variable X and $k = 1, 2, \ldots$, let $\mu_k = E(X^k)$, the kth moment of X. Suppose X has n possible values $x_1, \ldots, x_n$. Show that the n probabilities

$$p_i = P(X = x_i) \qquad (i = 1, \ldots, n)$$

are determined by the first $n - 1$ moments. [*Hint*: The vector $\mu = (1, \mu_1, \ldots, \mu_{n-1})$ is determined from the vector $p = (p_1, \ldots, p_n)$ as $\mu = pM$ for a suitable matrix M. Show that M has rank n, because if there were a linear combination of its columns which was identically zero, there would be a polynomial of degree $n-1$ with n roots. Deduce that M has an inverse M^{-1}, so that $p = \mu M^{-1}$.]

41. (*Hard.*) Suppose you toss a coin ten times and record the exact sequence of outcomes, e.g.,

H T H H T T H H T H .

Of course, many other sequences are possible. About how many times n would you have to repeat this ten toss experiment

a) to be 90% sure of seeing this particular sequence again in these n repetitions?

b) to be 90% sure of seeing at least one of the possible sequences twice in the n repetitions?

c) to be 90% sure of seeing every possible sequence at least once in the n repetitions?

d) to be 90% sure of seeing at least once every sequence in a set comprising exactly half of all possible outcomes, where the set is specified in advance.

e) Same as d), but for a set not specified in advance.