CHAPTER FOUR

DISCRETE PROBABILITY DISTRIBUTIONS

SECTION 4.1 Introduction

In Chapter 3 probability was defined and some of the basic tools used in working with probabilities were introduced. We now look at problems that can be put in a probabilistic framework. That is, by assessing the probabilities of certain events from actual past data, specific probability models that fit our problems can be used.

EXAMPLE 4.1 **Ophthalmology** Retinitis pigmentosa is a progressive ocular disease that in some cases eventually results in blindness. The three main genetic forms of the disease are the dominant mode, the recessive mode, and the sex-linked mode. Each mode has a different rate of progression, with the dominant mode being the slowest to progress and the sex-linked mode the fastest. Suppose a man does not have a clear idea of the prior history of the disease in his family but he does know that 1 of his 2 male children is affected, whereas his 1 female child is not affected. Can this information help identify the genetic type? ■■■

The **binomial distribution** can be applied to calculate the probability of this event occurring (1 out of 2 males affected, 0 out of 1 female affected) under each of the genetic modes mentioned, and these results can then be used to infer the most likely genetic mode. In fact, this distribution can be used to make an inference for any family where we know that k_1 out of n_1 male children are affected and k_2 out of n_2 female children are affected.

EXAMPLE 4.2 **Cancer** A second example of a commonly used probability model concerns a cancer scare in young children in Woburn, Massachusetts. A news story reported an "excessive" number of cancer deaths in young children in this town and speculated whether or not this high rate was due to the dumping of industrial wastes in the northeastern portion of town [1]. Suppose that 5 cases of leukemia were reported in a town where 1 would normally be expected. Is this difference sufficient evidence for concluding that there is an association between the industrial wastes and the cancer cases? ■■■

The **Poisson distribution** can be used to calculate the probability of five or more cases if typical national rates for cancer were present in this town. If this probability were sufficiently small, then we would conclude that there was an association; otherwise, we would conclude that a longer surveillance of the town was necessary before arriving at a conclusion.

In this chapter the general concept of a discrete random variable is introduced and the binomial and Poisson distributions are described in depth. This forms the basis for the discussion of hypothesis tests based on the binomial and Poisson distributions found in Chapters 7 and 10.

SECTION 4.2 Random Variables

In Chapter 3 we dealt with very specific events, such as the outcome of a tuberculin skin test or blood-pressure measurements taken on different members of a family. We now want to introduce ideas that will enable us to refer, in general terms, to different types of events having the *same probabilistic structure*. For this purpose the concept of a random variable is introduced.

DEFINITION 4.1

A **random variable** is a numerical quantity that takes different values with specified probabilities. ■

Two types of random variables are discussed in this text: discrete and continuous.

DEFINITION 4.2

A random variable for which there exists a discrete set of values with specified probabilities is a **discrete random variable**. ■

EXAMPLE 4.3

Otolaryngology Otitis media is a disease of the middle ear and is one of the most frequent reasons for visiting a doctor in the first 2 years of life other than a routine well-baby visit. Let X be the random variable that represents the number of episodes of otitis media in the first 2 years of life. Then X is a discrete random variable, which takes on the values 0, 1, 2, . . . ■■■

EXAMPLE 4.4

Hypertension Many new drugs have been introduced in the last decade to bring hypertension under control, that is, to reduce high blood pressure to normotensive levels. Suppose a physician agrees to use a new antihypertensive drug on a trial basis on the first 4 untreated hypertensives whom she encounters in her practice before deciding whether to adopt the drug for routine use. Let X = the number of patients out of 4 who are brought under control. Then X is a discrete random variable, which takes on the values 0, 1, 2, 3, 4. ■■■

DEFINITION 4.3

A random variable whose values form a continuum (i.e., have no gaps), such that ranges of values occur with specified probabilties, is a **continuous random variable**. ■

EXAMPLE 4.5

Environmental Health The possible health effects on workers exposed to low levels of radiation over long periods of time are an issue of public health interest. One problem in assessing this situation is how to measure the cumulative exposure of a worker. A study was performed at the Portsmouth Naval Shipyard, whereby each exposed worker wore a badge, or dosimeter, which measured annual radiation exposure in rem [2]. The cumulative exposure over a worker's lifetime could then be obtained by summing the yearly exposures. The cumulative lifetime exposure is a good example of a continuous random variable, since it varied in this study from 0.000 rem to 91.414 rem, which would be regarded as taking on an essentially infinite number of values. ■■■

SECTION 4.3 The Probability Mass Function for a Discrete Random Variable

The values taken by a discrete random variable and its associated probabilities can be expressed by a rule, or relationship, which is called a probability mass function.

DEFINITION 4.4

A **probability mass function** is a mathematical relationship, or rule, that assigns to any possible value r of a discrete random variable X the probability $Pr(X = r)$. This assignment is made for all values r that have positive probability. The probability mass function is sometimes also referred to as a **probability distribution**. ■

The probability mass function can be displayed in the form of a table giving the values and their associated probabilities and/or it can be expressed as a mathematical formula giving the probability of all possible values.

EXAMPLE 4.6

Hypertension Consider the situation in Example 4.4. Suppose that from previous experience with the drug, the drug company expects that for any clinical practice the probability that 0 patients out of 4 will be brought under control is .008, 1 patient out of 4 is .076, 2 patients out of 4 is .265, 3 patients out of 4 is .411, and all 4 patients is .240. This probability mass function, or probability distribution, is displayed in Table 4.1. ■■■

TABLE 4.1 Probability mass function for the hypertension-control example

$Pr(X = r)$	.008	.076	.265	.411	.240
r	0	1	2	3	4

Notice that for any probability mass function, the probability of any particular value must be between 0 and 1 and the sum of the probabilities of all values must exactly equal 1. Thus, $0 < Pr(X = r) \leq 1$, $\Sigma Pr(X = r) = 1$, where the summation is taken over all possible values that have positive probability.

EXAMPLE 4.7

Hypertension In Table 4.1, for any clinical practice, the probability that between 0 and 4 hypertensives are brought under control = 1; that is,

$$.008 + .076 + .265 + .411 + .240 = 1$$

■■■

4.3.1 Relationship of Probability Distributions to Sample Distributions

In Chapters 1 and 2 the concept of a **frequency distribution** in the context of a sample was discussed. It was described as a list of each value in the data set and a corresponding count of how frequently the values occur. If each count is divided by the total number of points in the sample, then the frequency distribution can be considered as a sample analogue to a probability distribution. In particular, a probability distribution can be thought of as a model based on an infinitely large sample, giving the fraction of data points in a sample that *should* be allocated to each specific value. Since the frequency distribution gives the actual proportion of points in a sample that correspond to specific values, the appropriateness of the model can be validated by comparing the observed sample frequency distribution to the probability distribution. The formal statistical procedure for making this comparison is called a **goodness-of-fit test**, which is discussed in Chapter 10.

EXAMPLE 4.8

Hypertension How can the probability mass function in Table 4.1 be used to see if the drug behaves with the same efficacy in actual practice as predicted by the drug company? The drug company might distribute the drug to 100 physicians and ask each of them to treat their first 4 untreated hypertensives with it. Each physician would then report his or her results to the drug company, and the combined results could be compared with the expected results in Table 4.1. For example, suppose that out of 100 physicians who agree to participate, 19 are able to bring all of their first 4 untreated hypertensives under control, 48 are able to bring 3 of the 4 hypertensives under control, 24 are able to bring 2 out of 4 under control, 9 are able to bring only 1 of 4 under control, and none of the physicians brings 0 out of 4 hypertensives under control. The sample-frequency distribution can be compared with the probability distribution given in Table 4.1. This comparison is shown in Table 4.2.

TABLE 4.2 Comparison of the sample-frequency distribution and the theoretical-probability distribution for the hypertension-control example

Number of hypertensives under control = r	Probability distribution $Pr(X = r)$	Frequency distribution
0	.008	.000 = 0/100
1	.076	.090 = 9/100
2	.265	.240 = 24/100
3	.411	.480 = 48/100
4	.240	.190 = 19/100

The distributions look reasonably similar. The role of statistical inference is to compare the two distributions to judge if the differences between the two can be attributed to chance or whether real differences exist between the drug's performance in actual clinical practice and expectations from previous drug-company experience. ■■■

A question often asked is: Where does a probability mass function come from? In some instances previous data can be obtained on the same type of random variable being studied and the probability mass function can be computed from these data. In other instances, previous data may not be available, but the probability mass function from some well-known distribution may be used to see how well it fits with some sample data. In fact, this approach was used in Table 4.2, where the probability mass function was derived from the binomial distribution and then compared with the frequency distribution from the sample of 100 physician practices.

SECTION 4.4 The Expected Value of a Discrete Random Variable

If a random variable has a large number of values with positive probability, then the probability mass function is not a useful summary measure. Indeed, we are faced with the same problem as in trying to summarize a sample by enumerating each data value.

Measures of location and spread can be developed for a random variable in much the same way as they were developed for samples. The analogue to the arithmetic mean $\bar{x}$ is referred to as the expected value of the random variable, or population mean, and is denoted by $E(X)$ or μ. The expected value represents the "average" value of the random variable. It is obtained by multiplying each possible value by its respective probability and summing over all the values that have positive (that is, non-zero) probability.

DEFINITION 4.5

The **expected value of a discrete random variable** is defined as

$$E(X) \equiv \mu = \sum_{i=1}^{k} x_i Pr(X = x_i)$$

where the x_i's are the values the random variable assumes with positive probability. ■

EXAMPLE 4.9 **Hypertension** Find the expected value for the random variable depicted in Table 4.1.

SOLUTION

$$E(X) = 0(.008) + 1(.076) + 2(.265) + 3(.411) + 4(.240) = 2.80$$

Thus, on average about 2.8 hypertensives would be expected to be brought under control for every 4 that are treated. ■■■

EXAMPLE 4.10 **Otolaryngology** Consider the random variable mentioned in Example 4.3 representing the number of episodes of otitis media in the first 2 years of life. Suppose this random variable has a probability mass function as given in Table 4.3.

TABLE 4.3 Probability mass function for the number of episodes of otitis media in the first 2 years of life

r	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

What is the expected number of episodes of otitis media in the first 2 years of life?

SOLUTION

$$E(X) = 0(.129) + 1(.264) + 2(.271) + 3(.185) + 4(.095) + 5(.039) + 6(.017) = 2.038$$

Thus, on the average a child would be expected to have 2 episodes of otitis media in the first 2 years of life. ■■■

In Example 4.8 the probability mass function for the random variable representing the number of previously untreated hypertensives brought under control was compared with the actual number of hypertensives brought under control in 100 clinical practices. In much the same way, the expected value of a random variable can be compared with the actual sample mean in a data set ($\bar{x}$).

EXAMPLE 4.11 **Hypertension** Compare the number of hypertensives brought under control in the 100 clinical practices ($\bar{x}$) with the expected number of hypertensives brought under control (μ).

SOLUTION From Table 4.2 we have

$$\bar{x} = [0(0) + 1(9) + 2(24) + 3(48) + 4(19)]/100 = 2.77$$

hypertensives controlled per clinical practice while $\mu = 2.80$. This agreement is rather good. The specific methods for comparing the observed average value and expected value of a random variable ($\bar{x}$ and μ) will be covered in the material on statistical inference in Chapter 7. Notice that $\bar{x}$ could be written in the form

$$\bar{x} = 0(0/100) + 1(9/100) + 2(24/100) + 3(48/100) + 4(19/100)$$

that is, as a weighted average of the number of hypertensives brought under control, where the weights are the observed probabilities. The expected value, in comparison, can be written as a similar weighted average, where the weights are the theoretical probabilities:

$$\mu = 0(.008) + 1(.076) + 2(.265) + 3(.411) + 4(.240)$$

Thus, the two quantities are actually obtained in the same way, one with weights given by the "observed" probabilities and the other with weights given by the "theoretical" probabilities.

■■■

SECTION 4.5 The Variance of a Discrete Random Variable

The analogue to the sample variance (s^2) for a random variable is called the variance of the random variable, or population variance, and is denoted by $Var(X)$. The variance represents the spread of all values that have positive probability relative to the expected value. In particular, the variance is obtained by multiplying the squared distance of each possible value from the expected value by its respective probability and summing over all the values that have positive probability.

DEFINITION 4.6

The **variance of a discrete random variable** denoted by X is defined by

$$Var(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 Pr(X = x_i)$$

where the x_i's are the values for which the random variable takes on positive probability. The **standard deviation of a random variable** X, denoted by $sd(X)$ or σ, is defined by the square root of its variance. ■

There is also a short form for the population variance, which is similar to the equation presented for the sample variance.

4.1 **A short form for the population variance** is given by

$$\sigma^2 = E(X - \mu)^2 = \sum_{i=1}^{k} x_i^2 Pr(X = x_i) - \mu^2$$

This can be obtained by expanding $(x_i - \mu)^2$ in the form $x_i^2 - 2x_i\mu + \mu^2$ and rewriting $Var(X)$ as

$$Var(X) = \sum_{i=1}^{k} (x_i^2 - 2\mu x_i + \mu^2) Pr(X = x_i)$$

$$= \sum_{i=1}^{k} x_i^2 Pr(X = x_i) + \sum_{i=1}^{k} (-2\mu) x_i Pr(X = x_i) + \sum_{i=1}^{k} \mu^2 Pr(X = x_i)$$

Since -2μ and μ^2 are constants, this expression can be rewritten in the form

$$Var(X) = \sum_{i=1}^{k} x_i^2 Pr(X = x_i) - 2\mu \sum_{i=1}^{k} x_i Pr(X = x_i) + \mu^2 \sum_{i=1}^{k} Pr(X = x_i)$$

Since, by definition, $\sum_{i=1}^{k} x_i Pr(X = x_i) = E(X) = \mu$, and $\sum_{i=1}^{k} Pr(X = x_i) = 1$, it follows that

$$Var(X) = \sum_{i=1}^{k} x_i^2 \, Pr(X = x_i) - 2\mu^2 + \mu^2 = \sum_{i=1}^{k} x_i^2 \, Pr(X = x_i) - \mu^2$$

EXAMPLE 4.12 **Otolaryngology** Compute the variance and standard deviation for the random variable depicted in Table 4.3.

SOLUTION We know from Example 4.10 that $\mu = 2.04$. Furthermore,

$$\begin{aligned}\sum_{i=1}^{k} x_i^2 \, Pr(X = x_i) &= 0^2(.129) + 1^2(.264) + 2^2(.271) + 3^2(.185) \\ &\quad + 4^2(.095) + 5^2(.039) + 6^2(.017) \\ &= 0(.129) + 1(.264) + 4(.271) + 9(.185) \\ &\quad + 16(.095) + 25(.039) + 36(.017) \\ &= 6.12\end{aligned}$$

Thus, $Var(X) = \sigma^2 = 6.12 - (2.038)^2 = 1.967$. The standard deviation of X is $\sigma = \sqrt{1.967} = 1.402$. ■■■

How can we get a feel for what the standard deviation of a random variable means? The following often-used principle is true for many, but not all, random variables:

4.2 **Approximately 95% of the probability mass falls within two standard deviations of the mean of a random variable.**

If 1.96σ is substituted for 2σ in equation 4.2, this statement holds exactly for normally distributed random variables and approximately for certain other random variables. Normally distributed random variables are discussed in detail in Chapter 5.

EXAMPLE 4.13 **Otolaryngology** Find a, b such that approximately 95% of infants will have between a and b episodes of otitis media in the first 2 years of life.

SOLUTION The random variable depicted in Table 4.3 has mean $(\mu) = 2.038$ and standard deviation $(\sigma) = 1.402$. The interval $\mu \pm 2\sigma$ is given by

$$2.038 \pm 2(1.402) = 2.038 \pm 2.805$$

or from -0.77 to 4.84. Since only positive integer values are possible for this random variable, the valid range is from $a = 0$ to $b = 4$ episodes. In Table 4.3 the probability of having ≤ 4 episodes is given as

$$.129 + .264 + .271 + .185 + .095 = .944$$ ■■■

The rule allows us to quickly summarize the range of values that have most of the probability mass for a random variable without specifying each individual value. In Chapter 6 the type of random variable for which **(4.2)** applies is specified more precisely.

SECTION 4.6 The Cumulative-Distribution Function of a Discrete Random Variable

Many random variables are displayed in tables or figures in terms of a cumulative-distribution function rather than a distribution of probabilities of individual values as in Table 4.1. The basic concept is to assign to each individual value the sum of probabilities of all values that are no larger than the value being considered. This function is defined as follows:

DEFINITION 4.7

The **cumulative-distribution function** of a discrete random variable X is denoted by $F(x)$ and is defined by $Pr(X \leq x)$. ■

EXAMPLE 4.14

Otolaryngology Compute the cumulative-distribution function for the otitis media random variable in Table 4.3 and display it graphically.

SOLUTION The cumulative-distribution function is given by

$$F(x) = 0 \quad \text{if} \quad x < 0$$
$$F(x) = .129 \quad \text{if} \quad 0 \leq x < 1$$
$$F(x) = .393 \quad \text{if} \quad 1 \leq x < 2$$
$$F(x) = .664 \quad \text{if} \quad 2 \leq x < 3$$
$$F(x) = .849 \quad \text{if} \quad 3 \leq x < 4$$
$$F(x) = .944 \quad \text{if} \quad 4 \leq x < 5$$
$$F(x) = .983 \quad \text{if} \quad 5 \leq x < 6$$
$$F(x) = 1.0 \quad \text{if} \quad x \geq 6$$

The function can be displayed as shown in Figure 4.1. ■■■

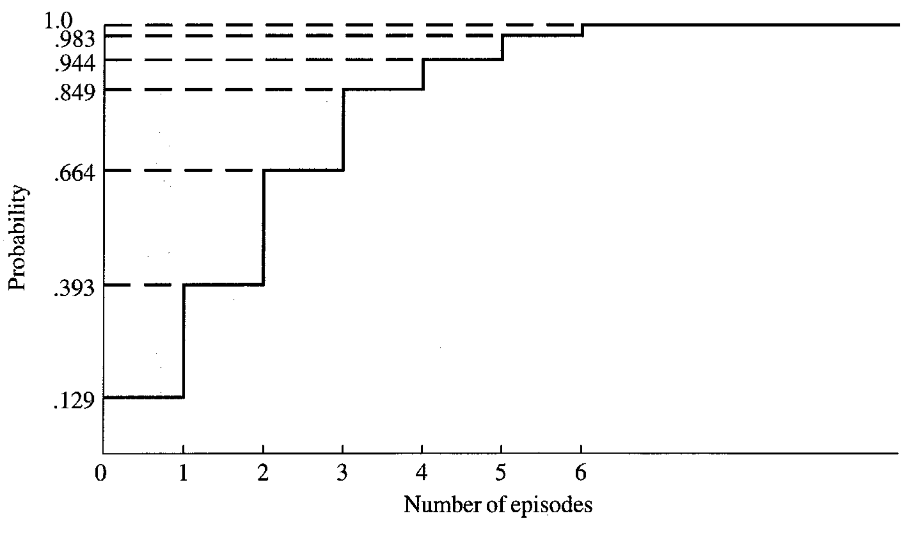


FIGURE 4.1 Cumulative-distribution function for the number of episodes of otitis media in the first 2 years of life

The cumulative distribution for a discrete random variable looks like a series of steps. The steps become smaller as the number of values increases, and the function approaches that of a smooth curve.

SECTION 4.7 Permutations and Combinations

In Sections 4.2 through 4.6 the concept of a discrete random variable was introduced in very general terms. In the remainder of this chapter, the focus is on some specific discrete random variables that occur frequently in medical and biological work. Consider the following example.

EXAMPLE 4.15

Infectious Disease One of the most common laboratory tests performed on any routine medical examination is a blood count. The two main aspects to a blood count are (1) counting the number of white blood cells (referred to as the "white count") and (2) differentiating white blood cells that do exist into five categories, namely, neutrophils, lymphocytes, monocytes, eosinophils, and basophils (referred to as the "differential"). Both the white count and the differential are extensively used in making clinical diagnoses. We will concentrate here on the differential, particularly on the distribution of the number of neutrophils k out of 100 white blood cells (which is the typical number counted). We will see that the number of neutrophils follows a binomial distribution. ■■■

To study the binomial distribution, **permutations** and **combinations**, important topics in probability, must first be understood.

EXAMPLE 4.16

Mental Health Suppose we identify 5 male subjects aged 50–59 with schizophrenia in a community, and we wish to match these subjects with normal controls of the same sex and age living in the same community. Suppose we wish to employ a **matched-pair design**, where each case is matched with a normal control of the same sex and age. Five psychologists are employed by the study, with each psychologist interviewing a single case and his matched control. If there are 10 eligible 50–59-year-old male controls in the community (labeled $A, B, \ldots, J$), then how many ways are there of choosing controls for the study if a control can never be used more than once?

SOLUTION

The first control can be any of $A, \ldots, J$ and thus can be chosen in 10 ways. Once the first control is chosen, he can no longer be selected as the second control; therefore, the second control can be chosen in 9 ways. Thus, the first two controls can be chosen in any one of $10 \times 9 = 90$ ways. Similarly, the third control can be chosen in any one of 8 ways, the fourth control in 7 ways, and the fifth control in 6 ways. In total, there are $10 \times 9 \times 8 \times 7 \times 6 =$ 30,240 ways of choosing the 5 controls. For example, one possible selection is *ACDFE*. This means that control A is matched to the first case, control C to the second case, and so on. The order of selection of the controls is important, since different psychologists may be assigned to interview each matched pair. Thus, the selection *ABCDE* is different from *CBAED*, even though the same group of controls is selected. ■■■

We can now ask the general question, how many ways can k objects be selected out of n where the order of selection matters? Note that the first object can be selected in any one of $n = (n + 1) - 1$ ways. Given that the first object has been selected, the second object can be selected in any one of $n - 1 = (n + 1) - 2$ ways; . . . ; the kth object can be selected in any one of $n - k + 1 = (n + 1) - k$ ways.

DEFINITION 4.8

The number of **permutations** of n things taken k at a time is

$$_nP_k = n(n-1) \times \cdots \times (n-k+1)$$

It represents the number of ways of selecting k items out of n, where the order of selection is important. ■

EXAMPLE 4.17

Mental Health Suppose there are 3 female schizophrenics aged 50–59 and 6 eligible controls living in the same community. How many ways are there of selecting three controls?

SOLUTION

To answer this question, consider the number of permutations of 6 things taken 3 at a time.

$$_6P_3 = 6 \times 5 \times 4 = 120$$

Thus, there are 120 ways of choosing the controls. For example, one way would be to match control A to case 1, control B to case 2, and control C to case 3 (i.e., ABC). Another way would be to match control F to case 1, control C to case 2, and control D to case 3 (i.e., FCD). The order of selection is important, since, for example, the selection ABC is different than the selection BCA. ■■■

In some instances we are interested in a special type of permutation: selecting n objects out of n, where the order of selection matters (i.e., ordering n objects). By the preceding principle,

$$_nP_n = n(n-1) \times \cdots \times [(n+1)-n] = n(n-1) \times \cdots \times 2 \times 1$$

The special symbol generally used for this quantity is $n!$, which is called n factorial and is defined as follows:

DEFINITION 4.9

$n! = n$ **factorial** is defined as $\quad n(n-1) \times \cdots \times 2 \times 1$ ■

EXAMPLE 4.18

Evaluate 5 factorial.

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$ ■■■

The quantity $0!$ has no intuitive meaning, but for consistency it will be defined as 1.

EXAMPLE 4.19

Mental Health Consider a somewhat different study design for the situation described in Example 4.16. Suppose an **unmatched study design,** whereby *all* cases and controls will be interviewed by the same psychologist, is used. If there are 10 eligible controls, then how many ways are there of choosing 5 controls for the study?

SOLUTION

In this case, since the same psychologist interviews all patients, what is important is which controls are selected, not the order of selection. Thus, the question is, how many ways can 5 out of 10 eligible controls be selected, where order is not important? Note that for each set of 5 controls (say A, B, C, D, E), there are $5 \times 4 \times 3 \times 2 \times 1 = 5!$ ways of ordering the controls among themselves (e.g., $ACBED$ and $DBCAE$ are two possible orders). Thus, the number of ways of selecting 5 out of 10 controls for the study without respect to order = (the number of ways of selecting 5 controls out of 10 where order is important)/5! = $_{10}P_5/5!$ = $(10 \times 9 \times 8 \times 7 \times 6)/120 = 30{,}240/120 = 252$ ways. Thus, $ABCDE$ and $CDFIJ$ are two possible selections. Also, $ABCDE$ and $BCADE$ are *not* counted twice. ■■■

The number of ways of selecting 5 objects out of 10 without respect to order is referred to as the number of **combinations** of 10 things taken 5 at a time and is denoted by ${}_{10}C_5$ or $\binom{10}{5} = 252$.

This discussion can be generalized to evaluate the number of combinations of n things taken k at a time. Note that for every selection of k distinct items out of n, there are $k(k-1) \times \cdots \times (2) \times (1) = k!$ ways of ordering the items among themselves. Thus, we have the following definition:

DEFINITION 4.10

The number of **combinations** of n things taken k at a time is

$$ {}_nC_k = \binom{n}{k} = \frac{n(n-1) \times \cdots \times (n-k+1)}{k!} $$ ■

By multiplying the numerator and denominator of ${}_nC_k$ by

$$ (n-k)! = (n-k)(n-k-1) \times \cdots \times 2 \times 1, $$

we obtain

$$ {}_nC_k = \frac{n(n-1) \times \cdots \times (n-k+1)}{k!} \times \frac{(n-k)(n-k-1) \times \cdots \times 2 \times 1}{(n-k)(n-k-1) \times \cdots \times 2 \times 1} $$

Note that the numerator $= n!$ and the denominator $= k! \times (n-k)!$. Thus, we have the following alternative definition:

DEFINITION 4.11

The number of **combinations** of n things taken k at a time is

$$ {}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} $$

It represents the number of ways of selecting k objects out of n where the order of selection does not matter. ■

EXAMPLE 4.20 Evaluate ${}_7C_3$.

$$ {}_7C_3 = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 7 \times 5 = 35 $$ ■■■

A special situation arises upon evaluating $\binom{n}{0}$. By definition, $\binom{n}{0} = n!/(0!n!)$, and 0! was defined as 1. Hence, $\binom{n}{0} = 1$ for any n.

Frequently, $\binom{n}{k}$ will need to be computed for $k = 0, 1, \ldots, n$, The combinatorials have the following symmetry property, which makes this calculation easier than it appears at first glance.

4.3 For any nonnegative integers n, k where $n \geq k$,

$$\binom{n}{k} = \binom{n}{n-k}$$

To see this, note from Definition 4.11 that

$$_nC_k = \frac{n!}{k!(n-k)!}$$

If $n - k$ is substituted for k in this expression, then we obtain

$$_nC_{n-k} = \frac{n!}{(n-k)![n-(n-k)]!} = \frac{n!}{(n-k)!k!} = {_nC_k}$$

Intuitively, this result makes sense, since $_nC_k$ represents the number of ways of selecting k objects out of n without regard to order. However, for every selection of k objects, we have also, in a sense, identified the other $n - k$ objects that were not selected. Thus, the number of ways of selecting k objects out of n without regard to order should be the same as the number of ways of selecting $n - k$ objects out of n without regard to order.

Hence, we need only evaluate combinatorials $\binom{n}{k}$ for the integers $k \leq n/2$. If $k > n/2$, then the relationship $\binom{n}{n-k} = \binom{n}{k}$ can be used.

EXAMPLE 4.21 Evaluate

$$\binom{7}{0}, \binom{7}{1}, \ldots, \binom{7}{7}$$

SOLUTION

$$\binom{7}{0} = 1 \quad \binom{7}{1} = 7 \quad \binom{7}{2} = \frac{7 \times 6}{2 \times 1} = 21 \quad \binom{7}{3} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$$

$$\binom{7}{4} = \binom{7}{3} = 35 \quad \binom{7}{5} = \binom{7}{2} = 21 \quad \binom{7}{6} = \binom{7}{1} = 7 \quad \binom{7}{7} = \binom{7}{0} = 1$$ ■■■

SECTION 4.8 The Binomial Distribution

All examples involving the binomial distribution have a common structure: a sample of n independent trials, each of which can have only two possible outcomes, which are denoted as "success" and "failure." Furthermore, the probability of a success at each trial is assumed to be some constant p, and hence the probability of a failure at each trial is $1 - p = q$. The term "success" is used in a general way, without any specific contextual meaning.

For Example 4.15, $n = 100$ and a "success" occurs when a cell is a neutrophil.

EXAMPLE 4.22 **Infectious Disease** Reconsider Example 4.15 with 5 cells rather than 100 and ask the more limited question, what is the probability that the second and fifth cells considered will be neutrophils and the remaining cells nonneutrophils given that the probability that any one cell is a neutrophil is .6?

SOLUTION If a neutrophil is denoted by an x and a nonneutrophil by an o, then the question being asked is, What is the probability of the outcome $oxoox = Pr(oxoox)$? Since the probabilities of success and failure are given respectively by .6 and .4, and the outcomes for different cells are presumed to be independent, then the probability is

$$q \times p \times q \times q \times p = p^2q^3 = (.6)^2(.4)^3$$ ■■■

EXAMPLE 4.23 **Infectious Disease** Now consider the more general question, What is the probability that any 2 cells out of 5 will be neutrophils?

SOLUTION The arrangement *oxoox* is only one of many possible orderings that result in 2 neutrophils. The 10 possible orderings are given in Table 4.4.

TABLE 4.4 Possible orderings for 2 neutrophils out of 5 cells

xxooo	*oxxoo*	*ooxox*
xoxoo	*oxoxo*	*oooxx*
xooxo	*oxoox*	
xooox	*ooxxo*	

In terms of combinations, the number of orderings = the number of ways of selecting 2 cells to be neutrophils out of 5 cells = ${}_5C_2 = (5 \times 4)/(2 \times 1) = 10$.

The probability of any of the orderings in Table 4.4 is the same as that for the ordering *oxoox*, namely, $(.6)^2(.4)^3$. Thus, the probability of obtaining 2 neutrophils in 5 cells is ${}_5C_2(.6)^2(.4)^3 = 10(.6)^2(.4)^3 = .230$. ■■■

Suppose the neutrophils problem is now considered more generally, with n trials rather than 5 trials, and the following question is asked: What is the probability of k successes (rather than 2 successes) in these n trials? The probability that the k successes will occur at k **specified** trials within the n trials and that the remaining trials will be failures is given by $p^k(1-p)^{n-k}$. To compute the probability of k successes in any of the n trials, this probability must be multiplied by the number of ways in which k trials for the successes and $n-k$ trials for the failures can be selected $= \binom{n}{k}$ (as was done in Table 4.4). Thus, the probability of k successes in n trials, or k neutrophils in n cells, is

$$\binom{n}{k}p^k(1-p)^{n-k} = \binom{n}{k}p^kq^{n-k}$$

4.4 The distribution of the number of successes in n statistically independent trials, where the probability of success on each trial is p, is known as the **binomial distribution** and has a probability mass function given by

$$Pr(X = k) = \binom{n}{k}p^kq^{n-k}, \qquad k = 0, 1, \ldots, n$$

EXAMPLE 4.24 What is the probability of obtaining 2 boys out of 5 children if the probability of a boy is .51 at each birth and the sexes of successive children are considered independent random variables?

SOLUTION Use a binomial distribution with $n = 5, p = .51, k = 2$. Compute

$$Pr(X = 2) = {}_5C_2(.51)^2(.49)^3 = \frac{5 \times 4}{2 \times 1}(.51)^2(.49)^3$$
$$= 10(.51)^2(.49)^3 = .306$$

■■■

4.8.1 Using Binomial Tables

Frequently, a number of binomial probabilities will need to be evaluated for the same n and p, which would be tedious if each probability had to be calculated from **(4.4)**. Instead, for small n ($n \leq 20$) and selected values of p, refer to Table 1 in the Appendix, where the individual binomial probabilities are calculated. In this table, the number of trials (n) is provided in the first column, the number of successes (k) out of the n trials is given in the second column, and the probability of success for an individual trial (p) is given in the first row. Binomial probabilities are provided for $n = 2, 3, \ldots, 20$, $p = .05, .10, \ldots, .50$.

EXAMPLE 4.25 **Infectious Disease** Evaluate the probability of 2 lymphocytes out of 10 white blood cells if the probability that any one cell is a lymphocyte is .2.

SOLUTION Refer to Table 1 with $n = 10, k = 2, p = .20$. The appropriate probability, given in the $k = 2$ row and $p = .20$ column under $n = 10$, is .3020. ■■■

EXAMPLE 4.26 **Pulmonary Disease** An investigator notices that children develop chronic bronchitis in the first year of life in 3 out of 20 households where both parents are chronic bronchitics, as compared with the national incidence rate of chronic bronchitis, which is 5% in the first year of life. Is this difference "real" or can it be attributed to chance? Specifically, how likely are infants in at least 3 out of 20 households to develop chronic bronchitis if the probability of developing disease in any one household is .05?

SOLUTION Suppose the underlying rate of disease in the offspring is .05. Under this assumption, the number of households where the infants develop chronic bronchitis will follow a binomial distribution with parameters $n = 20, p = .05$. Thus, the probability of observing k cases out of 20 with disease is given by

$$\binom{20}{k}(.05)^k(.95)^{20-k}, \qquad k = 0, 1, \ldots, 20$$

The question is, What is the probability of observing at least 3 cases? The answer is

$$Pr(X \geq 3) = \sum_{k=3}^{20}\binom{20}{k}(.05)^k(.95)^{20-k} = 1 - \sum_{k=0}^{2}\binom{20}{k}(.05)^k(.95)^{20-k}$$

These 3 probabilities in the sum can be evaluated using the binomial table (Table 1). Refer to $n = 20, p = .05$ and note that $Pr(X = 0) = .3585$, $Pr(X = 1) = .3774$, $Pr(X = 2) = .1887$. Thus,

$$Pr(X \geq 3) = 1 - (.3585 + .3774 + .1887) = .0754$$

Thus, $X \geq 3$ is an unusual event, but not very unusual. If 3 infants out of 20 were to develop the disease, it would be difficult to judge whether the familial aggregation was real until a larger sample was available. ■■■

One question that arises is how to use the binomial tables if the probability of success on an individual trial (p) is greater than .5. Recall that

$$\binom{n}{k} = \binom{n}{n-k}$$

and let X be a binomial random variable with parameters n and p, and Y be a binomial random variable with parameters n and $q = 1 - p$. Then **(4.4)** can be rewritten as

4.5

$$Pr(X = k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{n-k} q^{n-k} p^k = Pr(Y = n - k)$$

In words, the probability of obtaining k successes for a binomial random variable X with parameters n and p is the same as the probability of obtaining $n - k$ successes for a binomial random variable Y with parameters n and q. Clearly, if $p > .5$, then $q = 1 - p < .5$, and Table 1 can be used with sample size n, referring to the $n - k$ row and the q column to obtain the appropriate probability.

EXAMPLE 4.27 **Infectious Disease** Evaluate the probabilities of obtaining k neutrophils out of 5 cells for $k = 0, 1, 2, 3, 4, 5$, where the probability that any one cell is a neutrophil is .6.

SOLUTION Since $p > .5$, refer to the random variable Y with parameters $n = 5$, $p = 1 - .6 = .4$.

$$Pr(X = 0) = \binom{5}{0}(.6)^0(.4)^5 = \binom{5}{5}(.4)^5(.6)^0 = Pr(Y = 5) = .0102$$

upon referring to the $k = 5$ row and $p = .40$ column under $n = 5$. Similarly,

$Pr(X = 1) = Pr(Y = 4) = .0768$ upon referring to the 4 row and .40 column under $n = 5$

$Pr(X = 2) = Pr(Y = 3) = .2304$ upon referring to the 3 row and .40 column under $n = 5$

$Pr(X = 3) = Pr(Y = 2) = .3456$ upon referring to the 2 row and .40 column under $n = 5$

$Pr(X = 4) = Pr(Y = 1) = .2592$ upon referring to the 1 row and .40 column under $n = 5$

$Pr(X = 5) = Pr(Y = 0) = .0778$ upon referring to the 0 row and .40 column under $n = 5$

■■■

4.8.2 Recursion Rule for Binomial Probabilities

In many instances we will want to evaluate binomial probabilities for $n > 20$ and/or for values of p not given in Table 1 of the Appendix. For sufficiently large n, the normal distribution can be used to approximate the binomial distribution, and tables of the normal distribution can be used to evaluate binomial probabilities. This procedure is usually less tedious than evaluating binomial probabilities directly using **(4.4)** and is studied in detail in Chapter 5. Alternatively, if the sample size is not large enough to use the normal approximation and if the value of p is not in Table 1, then a recursion rule can be used to evaluate binomial probabilities. This rule is particularly useful in evaluating many binomial probabilities for the same n and p. Using the recursion rule, it is easy to evaluate $Pr(X = k + 1)$ once $Pr(X = k)$ is known. Thus, once the probability of 0 successes has been computed, the probability of 1 success, 2 successes, and so forth can easily be computed without computing any combinatorials. The recursion rule is given as follows:

4.6 **Recursion Rule for Binomial Probabilities**

$$Pr(X = k + 1) = [(n - k)/(k + 1)] \times (p/q) \times Pr(X = k), \quad k = 0, 1, \ldots, n - 1$$

To see this, remember from **(4.4)** that

$$Pr(X = k + 1) = \binom{n}{k + 1} p^{k+1} q^{n-(k+1)} = \frac{n!}{(k + 1)!(n - k - 1)!} p^{k+1} q^{n-k-1}$$

$$Pr(X = k) = \frac{n!}{k!(n - k)!} p^k q^{n-k}$$

Divide $Pr(X = k + 1)$ by $Pr(X = k)$ to obtain

$$\frac{Pr(X = k + 1)}{Pr(X = k)} = \frac{\{n!/[(k + 1)!(n - k - 1)!]\} p^{k+1} q^{n-k-1}}{\{n!/[k!(n - k)!]\} p^k q^{n-k}}$$

$$= \frac{k!}{(k + 1)!} \times \frac{(n - k)!}{(n - k - 1)!} \times \frac{p}{q}$$

However, since $k!/(k + 1)! = 1/(k + 1)$ and $(n - k)!/(n - k - 1)! = n - k$, it follows that

$$\frac{Pr(X = k + 1)}{Pr(X = k)} = \frac{1}{k + 1} \times (n - k) \times \frac{p}{q}$$

Upon multiplying both sides of the equation by $Pr(X = k)$, we have

$$Pr(X = k + 1) = \frac{n - k}{k + 1} \times \frac{p}{q} \times Pr(X = k)$$

EXAMPLE 4.28 **Infectious Disease** Suppose that a group of 100 males aged 60–64 received a new flu vaccine in 1986 and that 5 of them died within the next year. Is this event unusual or can this death rate be expected for people of this age-sex group? Specifically, how likely are at least 5 out of 100 60–64-year-old males who receive a flu vaccine to die in the next year?

SOLUTION We first find the expected annual death rate in 60–64-year-old males. From a 1986 U.S. life table, we find that 60–64-year-old men have an approximate probability of death in the next year of .020 [3]. Thus, from the binomial distribution the probability that k out of 100 men will die during the next year is given by $\binom{100}{k}(.02)^k(.98)^{100-k}$. We want to know if 5 deaths in a sample of 100 men is an "unusual" event. One criterion for this evaluation might be to find the probability of getting at least 5 deaths in this group = $Pr(X \geq 5)$ given that the probability of death for an individual man is .02. This probability can be expressed as

$$\sum_{k=5}^{100} \binom{100}{k}(.02)^k(.98)^{100-k}$$

Because this sum of 96 probabilities is tedious to compute, we instead compute

$$Pr(X < 5) = \sum_{k=0}^{4} \binom{100}{k}(.02)^k(.98)^{100-k}$$

and then evaluate $Pr(X \geq 5) = 1 - Pr(X < 5)$. The binomial tables cannot be used because $n > 20$. Therefore, the sum of 5 binomial probabilities is evaluated using the recursion rule.

$$Pr(X = 0) = \binom{100}{0}(.02)^0(.98)^{100} = (.98)^{100} = .13262$$

$$Pr(X = 1) = \left(\frac{100 - 0}{0 + 1}\right)\left(\frac{.02}{.98}\right)(.13262) = .27065$$

$$Pr(X = 2) = \left(\frac{99}{2}\right)\left(\frac{.02}{.98}\right)(.27065) = .27341$$

$$Pr(X = 3) = \left(\frac{98}{3}\right)\left(\frac{.02}{.98}\right)(.27341) = .18228$$

$$Pr(X = 4) = \left(\frac{97}{4}\right)\left(\frac{.02}{.98}\right)(.18228) = .09021$$

Hence,

$$Pr(X < 5) = .13262 + .27065 + .27341 + .18228 + .09021 = .94917$$

and

$$Pr(X \geq 5) = 1 - Pr(X < 5) = .051$$

Thus, 5 deaths in 100 is a slightly unusual, but not a very unusual, event. If there were 10 deaths rather than 5, then using the same approach,

$$Pr(X \geq 10) = 1 - Pr(X < 10) < .001$$

which is very unlikely and would probably be grounds for halting the use of the vaccine in the absence of any other evidence. ■■■

SECTION 4.9 Expected Value and Variance of the Binomial Distribution

The expected value and variance of the binomial distribution are important both in terms of our general knowledge about the binomial distribution and for our later work on estimation and hypothesis testing. From Definition 4.5 we know that the general formula for the expected value of a discrete random variable is

$$E(X) = \sum_{i=1}^{k} x_i Pr(X = x_i)$$

Explicit derivations of mean and variance

[edit]

We derive these quantities from first principles. Certain particular sums occur in these two derivations. We rearrange the sums and terms so that sums solely over complete binomial probability mass functions (pmf) arise, which are always unity

$$\sum_{k=0}^{n} \Pr(X=k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Mean

[edit]

We apply the definition of the expected value of a discrete random variable to the binomial distribution

$$\operatorname{E}(X) = \sum_k x_k \cdot \Pr(x_k) = \sum_{k=0}^{n} k \cdot \Pr(X=k) = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k}.$$

The first term of the series (with index $k = 0$) has value 0 since the first factor, k, is zero. It may thus be discarded, i.e. we can change the lower limit to: $k = 1$

$$\operatorname{E}(X) = \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \cdot \frac{n \cdot (n-1)!}{k \cdot (k-1)!(n-k)!} \cdot p \cdot p^{k-1} (1-p)^{n-k}.$$

We've pulled factors of n and k out of the factorials, and one power of p has been split off. We are preparing to redefine the indices.

$$\operatorname{E}(X) = np \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k}$$

We rename $m = n - 1$ and $s = k - 1$. The value of the sum is not changed by this, but it now becomes readily recognizable

$$\operatorname{E}(X) = np \cdot \sum_{s=0}^{m} \frac{(m)!}{(s)!(m-s)!} p^s (1-p)^{m-s} = np \cdot \sum_{s=0}^{m} \binom{m}{s} p^s (1-p)^{m-s}.$$

The ensuing sum is a sum over a complete binomial pmf (of one order lower than the initial sum, as it happens). Thus

$$\operatorname{E}(X) = np \cdot 1 = np.$$

Variance

[edit]

It can be shown that the variance is equal to (see: variance, 10. Computational formula for variance):

$$\operatorname{Var}(X) = \operatorname{E}(X^2) - (\operatorname{E}(X))^2.$$

In using this formula we see that we now also need the expected value of X^2, which is

$$\operatorname{E}(X^2) = \sum_{k=0}^{n} k^2 \cdot \Pr(X=k) = \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} p^k (1-p)^{n-k}.$$

We can use our experience gained above in deriving the mean. We know how to process one factor of k. This gets us as far as

$$\operatorname{E}(X^2) = np \cdot \sum_{s=0}^{m} k \cdot \binom{m}{s} p^s (1-p)^{m-s} = np \cdot \sum_{s=0}^{m} (s+1) \cdot \binom{m}{s} p^s (1-p)^{m-s}$$

(again, with $m = n - 1$ and $s = k - 1$). We split the sum into two separate sums and we recognize each one

$$\operatorname{E}(X^2) = np \cdot \left(\sum_{s=0}^{m} s \cdot \binom{m}{s} p^s (1-p)^{m-s} + \sum_{s=0}^{m} 1 \cdot \binom{m}{s} p^s (1-p)^{m-s} \right).$$

The first sum is identical in form to the one we calculated in the Mean (above). It sums to mp. The second sum is unity.

$$\operatorname{E}(X^2) = np \cdot (mp + 1) = np((n-1)p + 1) = np(np - p + 1).$$

Using this result in the expression for the variance, along with the Mean ($\operatorname{E}(X) = np$), we get

$$\operatorname{Var}(X) = \operatorname{E}(X^2) - (\operatorname{E}(X))^2 = np(np - p + 1) - (np)^2 = np(1-p).$$

In the special case of a binomial distribution, the only values that take on positive probability are 0, 1, 2, . . . , n, and these values occur with probabilities

$$\binom{n}{0}p^0q^n, \binom{n}{1}p^1q^{n-1}, \ldots$$

Thus,
$$E(X) = \sum_{k=0}^{n} k\binom{n}{k}p^kq^{n-k}$$

It can be shown that this summation reduces to the simple expression np. Similarly, using Definition 4.6, we can show that

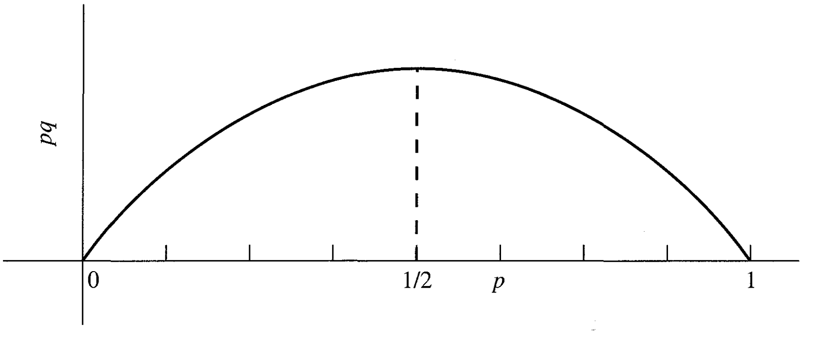
$$Var(X) = \sum_{k=0}^{n} (k - np)^2\binom{n}{k}p^kq^{n-k} = npq$$

which leads directly to the following result:

4.7 The **expected value and variance of a binomial distribution** are np and npq, respectively.

These results make good sense, since the expected number of successes in n trials is simply the probability of success on one trial multiplied by n, which equals np. Furthermore, for a given number of trials n, the binomial distribution has the highest variance when $p = \frac{1}{2}$, a shown in Figure 4.2. The variance of the distribution decreases as p moves away from $\frac{1}{2}$ in either direction, becoming 0 when $p = 0$ or 1. This result makes sense, since when $p = 0$ there must be 0 successes in n trials and when $p = 1$ there must be n successes in n trials, and there is no variability in either instance. Furthermore, when p is near 0 or near 1, the distribution of the number of successes is clustered near 0 and n, respectively, and there is comparatively little variability as compared with the situation when $p = \frac{1}{2}$. This point is depicted in Figure 4.3.

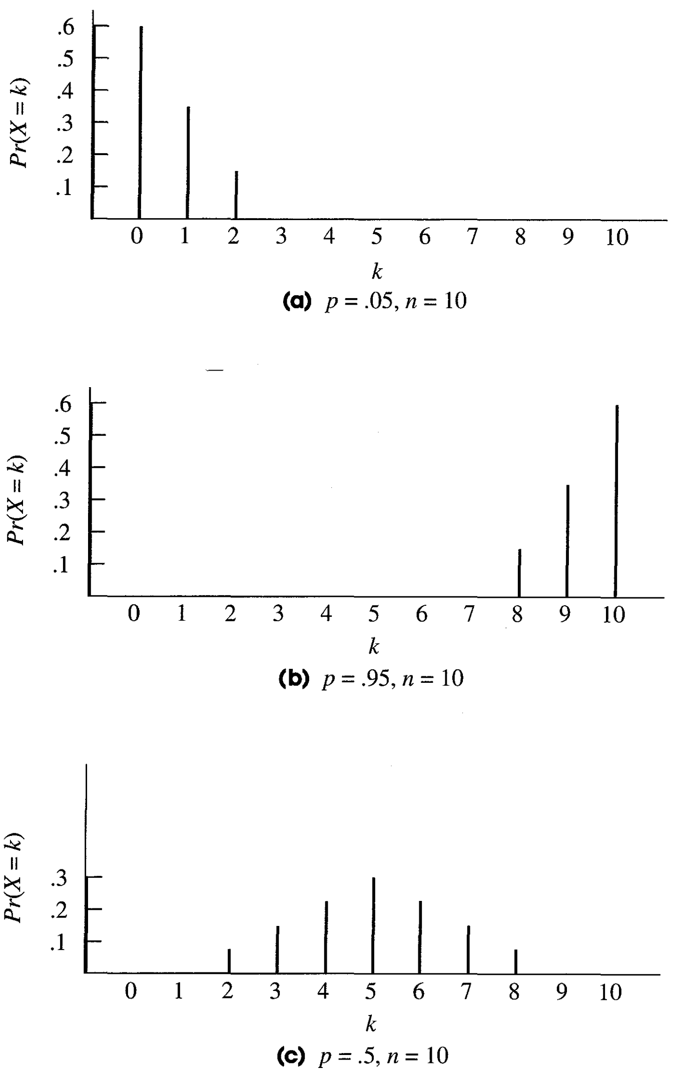
FIGURE 4.2
Plot of pq versus p



SECTION 4.10 The Poisson Distribution

The Poisson distribution is perhaps the second most frequently used discrete distribution after the binomial distribution. This distribution is usually associated with rare events.

FIGURE 4.3
The binomial distribution for various values of p when $n = 10$



EXAMPLE 4.29

Infectious Disease Consider the distribution of the number of deaths attributed to typhoid fever over a long period of time, for example, 1 year. Assuming that the probability of a new death from typhoid fever in any one day is very small and that the number of cases reported in any two distinct periods of time are independent random variables, then the number of deaths over a 1-year period will follow a Poisson distribution. ■■■

EXAMPLE 4.30

Bacteriology The preceding example concerns a rare event occurring over time. Rare events can also be considered not over time but on a surface area, such as the distribution of the number of bacterial colonies growing on an agar plate. Suppose we have a 100-cm^2 agar plate and that the probability of finding any bacterial colonies at any 1 point a (or more precisely in a small area around a) is very small and that the events of finding bacterial colonies at any 2 points a_1, a_2 are independent. The number of bacterial colonies over the entire agar plate will follow a Poisson distribution. ■■■

Consider Example 4.29. Ask the question, What is the distribution of the number of deaths due to typhoid fever from time 0 to time t (where t is some long period of time, such as 1 year or 20 years)?

Three assumptions must be made about the incidence of the disease. Consider any general *small* subinterval of the time period t, denoted by Δt.

ASSUMPTION 4.1 Assume that

(a) The probability of observing 1 death is directly proportional to the length of the time interval Δt. That is, $Pr(1 \text{ death}) \approx \lambda\, \Delta t$ for some constant λ.

(b) The probability of observing 0 deaths over Δt is approximately $1 - \lambda\, \Delta t$.

(c) The probability of observing more than 1 death over this time interval is essentially 0. ■

ASSUMPTION 4.2 **Stationarity** Assume that the number of deaths per unit time is the same throughout the entire time interval t. Thus, an increase in the incidence of the disease as time goes on within the time period t would violate this assumption. Note that t should not be overly long, since this assumption is less likely to hold as t increases. ■

ASSUMPTION 4.3 **Independence** If a death occurs within one time subinterval, it has no bearing on the probability of death in the next time subinterval. This assumption would be violated in an epidemic situation, because if a new case of disease occurs, then subsequent deaths are likely to build up over a short period of time until after the epidemic subsides. ■

Based on these assumptions, the Poisson probability distribution can be derived:

4.8 The probability of k events occurring in a time period t for a Poisson random variable with parameter λ is

$$Pr(X = k) = e^{-\mu}\mu^k/k!, \qquad k = 0, 1, 2, \ldots$$

where $\mu = \lambda t$ and e is approximately 2.71828.

Thus, the Poisson distribution depends on one parameter $\mu = \lambda t$. Note that the parameter λ represents the *expected number of events per unit time*, whereas the parameter μ represents the *expected number of events over the time period* t. One important difference between the Poisson distribution and the binomial distribution concerns the numbers of trials and events. For a binomial distribution there are a finite number of trials n, and the number of events can be no larger than n. For a Poisson distribution the number of trials is essentially infinite and the number of events (or number of deaths) can be indefinitely large, although the probability of k events will get very small as k gets large.

EXAMPLE 4.31 **Infectious Disease** Consider the typhoid-fever example. Suppose the number of deaths attributable to typhoid fever over a 1-year period is Poisson with parameter $\mu = 4.6$. What is the probability distribution of the number of deaths over a 6-month period? a 3-month period?

SOLUTION Let X = the number of deaths in 6 months. Since $\mu = 4.6$, $t = 1$, it follows that $\lambda = 4.6$. For a 6-month period we have that $\lambda = 4.6$, $t = .5$. Thus, $\mu = \lambda t = 2.3$. Therefore,

$$Pr(X = 0) = e^{-2.3} = .100$$

$$Pr(X = 1) = \frac{2.3}{1!}e^{-2.3} = .231$$

$$Pr(X = 2) = \frac{2.3^2}{2!}e^{-2.3} = .265$$

$$Pr(X = 3) = \frac{2.3^3}{3!}e^{-2.3} = .203$$

$$Pr(X = 4) = \frac{2.3^4}{4!}e^{-2.3} = .117$$

$$Pr(X = 5) = \frac{2.3^5}{5!}e^{-2.3} = .054$$

$$Pr(X \geq 6) = 1 - (.100 + .231 + .265 + .203 + .117 + .054) = .030$$

Let Y = the number of deaths in 3 months. For a 3-month period we have that $\lambda = 4.6$, $t = .25$, $\mu = \lambda t = 1.15$. Therefore,

$$Pr(Y = 0) = e^{-1.15} = .317$$

$$Pr(Y = 1) = \frac{1.15}{1!}e^{-1.15} = .364$$

$$Pr(Y = 2) = \frac{1.15^2}{2!}e^{-1.15} = .209$$

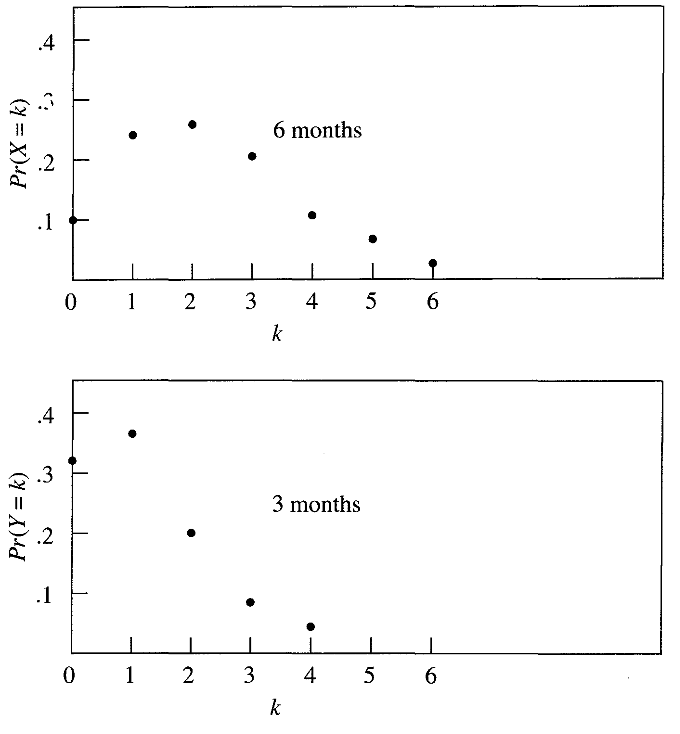
$$Pr(Y = 3) = \frac{1.15^3}{3!}e^{-1.15} = .080$$

$$Pr(Y \geq 4) = 1 - (.317 + .364 + .209 + .080) = .030$$

These distributions are plotted in Figure 4.4. Note that the distribution tends to become more symmetric as the time interval increases or, more specifically, as μ increases. ■■■

The Poisson distribution can also be applied to Example 4.30, where the distribution of the number of bacterial colonies in an agar plate of area A is discussed. Assuming that the probability of finding 1 colony in an area of size ΔA at any point on the plate is $\lambda \Delta A$ for some λ and that the number of bacterial colonies found at 2 different points of the plate are independent random variables, then the probability of finding k bacterial colonies in an area of size A is given by $e^{-\mu}\mu^k/k!$, where $\mu = \lambda A$.

FIGURE 4.4 Distribution of the number of deaths attributable to typhoid fever over various time intervals



EXAMPLE 4.32 **Bacteriology** If $A = 100 \text{ cm}^2$, $\lambda = .02$, calculate the probability distribution of the number of bacterial colonies.

SOLUTION We have that $\mu = \lambda A = 100(.02) = 2$. Let X = the number of colonies.

$$Pr(X = 0) = e^{-2} = .135$$

$$Pr(X = 1) = e^{-2}2^1/1! = 2e^{-2} = .271$$

$$Pr(X = 2) = e^{-2}2^2/2! = 2e^{-2} = .271$$

$$Pr(X = 3) = e^{-2}2^3/3! = \frac{4}{3}e^{-2} = .180$$

$$Pr(X = 4) = e^{-2}2^4/4! = \frac{2}{3}e^{-2} = .090$$

$$Pr(X \geq 5) = 1 - (.135 + .271 + .271 + .180 + .090) = .053$$

Clearly, the larger λ is, the more bacterial colonies we would expect to find. ■■■

SECTION 4.11 Computation of Poisson Probabilities

4.11.1 Using Poisson Tables

A number of Poisson probabilities for the same parameter μ often need to be evaluated. This task would be tedious if **(4.8)** had to be applied repeatedly. Instead, for $\mu \leq 20$

refer to Table 2 in the Appendix, in which individual Poisson probabilities are specifically calculated. In this table the Poisson parameter μ is given in the first row, the number of events (k) is given in the first column, and the corresponding Poisson probability is given in the k row and μ column.

EXAMPLE 4.33 Compute the probability of obtaining at least 5 events for a Poisson distribution with parameter $\mu = 3$.

SOLUTION Refer to Table 2 under the 3.0 column. Let X = the number of events.

$$Pr(X = 0) = .0498$$

$$Pr(X = 1) = .1494$$

$$Pr(X = 2) = .2240$$

$$Pr(X = 3) = .2240$$

$$Pr(X = 4) = .1680$$

Thus,

$$\begin{aligned} Pr(X \geq 5) &= 1 - Pr(X \leq 4) \\ &= 1 - (.0498 + .1494 + .2240 + .2240 + .1680) \\ &= 1 - .8152 = .1848 \end{aligned}$$

■■■

4.11.2 Recursion Rule for Poisson Probabilities

In many instances we will want to evaluate a collection of Poisson probabilities for the same μ, but μ will not be given in Table 2 of the Appendix. For large μ ($\mu \geq 10$), a normal approximation, as given in Chapter 5, can be used. Otherwise, the following recursion rule, which is similar to that given for binomial probabilities, can be used:

4.9

Recursion Rule for Poisson Probabilities

If $Pr(X = k)$ is the Poisson probability of observing k events with underlying parameter μ, then

$$Pr(X = k + 1) = [\mu/(k + 1)]Pr(X = k)$$

EXAMPLE 4.34 **Infectious Disease** Apply the recursion rule to the distribution of deaths due to typhoid fever over a 3-month period given in Example 4.31.

SOLUTION First, compute the probability of 0 deaths = $Pr(Y = 0) = e^{-1.15} = .3166$. Then,

$$Pr(Y = 1) = (1.15/1)Pr(Y = 0) = 1.15(.3166) = .3641$$

$$Pr(Y = 2) = (1.15/2)Pr(Y = 1) = (.575)(.3641) = .2094$$

$$Pr(Y = 3) = (1.15/3)Pr(Y = 2) = (1.15/3)(.2094) = .0803$$

■■■

SECTION 4.12 Expected Value and Variance of the Poisson Distribution

In many instances we cannot predict whether the assumptions for the Poisson distribution given in Section 4.10 are satisfied. Fortunately, the relationship between the expected value and variance of the Poisson distribution provides an important guideline that helps identify random variables that follow this distribution. This relationship can be stated as follows:

4.10 For a Poisson distribution with parameter μ, the mean and variance are both equal to μ.

This fact is useful to know, since if we have a data set from a discrete distribution where the *mean and variance are about the same*, then we can preliminarily identify it as a Poisson distribution and use various tests to confirm this hypothesis.

EXAMPLE 4.35 **Infectious Disease** The number of deaths attributable to polio during the years 1968–1976 are given in Table 4.5 [4, 5]. Comment on the applicability of the Poisson distribution to this data set.

SOLUTION The sample mean and variance of the annual number of deaths due to polio during the period 1968–1976 are 11.3 and 51.5, respectively. The Poisson distribution clearly will not fit well here, since the variance is 4.5 times as large as the mean. The larger variance is probably due to the clustering of polio deaths at certain times and geographical locations, which leads to a violation of both the independence assumption and the assumption of constant incidence over time. ■■■

TABLE 4.5 Number of deaths attributable to polio during the years 1968–1976

Year	1968	1969	1970	1971	1972	1973	1974	1975	1976
Number of deaths	24	13	7	18	2	10	3	9	16

Suppose we are studying a rare event phenomenon and wish to apply the Poisson distribution. A question that often arises is how to estimate the parameter μ of the Poisson distribution in this context. Since the expected value of the Poisson distribution is μ, μ can be estimated by the observed mean number of events, if such data are available. If the data are not available, other data sources can be used to estimate μ.

EXAMPLE 4.36 **Occupational Health** A public health issue arose concerning the possible carcinogenic potential of food ingredients containing ethylene dibromide (EDB). In some instances foods were removed from public consumption if they were shown to have excessive quantities of EDB. A study was previously performed looking at the mortality experience of 161 white male employees of two plants in Texas and Michigan who were exposed to EDB over the time period 1940–1975 [6]. Seven deaths due to cancer were observed among these employees. For this time period, 5.8 cancer deaths were expected as calculated from overall mortality rates for U.S. white males. Assess if the observed number of cancer deaths was excessive in this group.

SOLUTION Estimate the parameter μ from the expected number of cancer deaths from U.S. white male mortality rates; that is, $\mu = 5.8$. Then calculate $Pr(X \geqslant 7)$, where X is a Poisson random variable with parameter 5.8. Use the relationship

$$Pr(X \geqslant 7) = 1 - Pr(X \leqslant 6)$$

Since 5.8 is not in Table 2 of the Appendix, use the recursion rule.

$$Pr(X = 0) = \frac{e^{-5.8}(5.8)^0}{0!} = e^{-5.8} = .0030$$

$$Pr(X = 1) = \frac{5.8}{1} \times .0030 = .0176$$

$$Pr(X = 2) = \frac{5.8}{2} \times .0176 = .0509$$

$$Pr(X = 3) = \frac{5.8}{3} \times .0509 = .0985$$

$$Pr(X = 4) = \frac{5.8}{4} \times .0985 = .1428$$

$$Pr(X = 5) = \frac{5.8}{5} \times .1428 = .1656$$

$$Pr(X = 6) = \frac{5.8}{6} \times .1656 = .1601$$

Thus,

$$Pr(X \geqslant 7) = 1 - Pr(X \leqslant 6)$$

$$= 1 - (.0030 + \cdots + .1601) = 1 - .6384 = .362$$

Clearly, the observed number of cancer deaths is not excessive in this group. ■■■

SECTION 4.13 Poisson Approximation to the Binomial Distribution

As was seen in the preceding section, the Poisson distribution appears to fit well in some applications. Another important use for the Poisson distribution is as an approximation to the binomial distribution. Consider the binomial distribution for large n and small p. The mean of this distribution is given by np and the variance by npq. Note that $q \approx$ (is approximately equal to) 1 for small p, and thus $npq \approx np$. Therefore, the mean and variance of the binomial distribution are almost equal in this case, which suggests the following rule:

4.11 **Poisson Approximation to the Binomial Distribution**

The binomial distribution with large n and small p can be accurately approximated by a Poisson distribution with parameter $\mu = np$.

The rationale for using this approximation is that the Poisson distribution is easier to work with than the binomial distribution. The binomial distribution involves expressions such as $\binom{n}{k}$ and $(1 - p)^{n-k}$, which are cumbersome for large n.

EXAMPLE 4.37 **Cancer, Genetics** Suppose we are interested in the genetic susceptibility to breast cancer. We find that 4 out of 1000 women aged 40–49 whose mothers have had breast cancer develop breast cancer over the next year of life. We would expect from large population studies that 1 in 1000 women of this age group will develop a new case of the disease over this period of time. How unusual is this event?

SOLUTION The exact binomial probability could be computed by letting $n = 1000$, $p = 1/1000$. Hence,

$$\begin{aligned} Pr(X \geq 4) &= 1 - Pr(X \leq 3) \\ &= 1 - \left[\binom{1000}{0}(.001)^0(.999)^{1000} + \binom{1000}{1}(.001)^1(.999)^{999} \right. \\ &\quad \left. + \binom{1000}{2}(.001)^2(.999)^{998} + \binom{1000}{3}(.001)^3(.999)^{997}\right] \end{aligned}$$

Instead, use the Poisson approximation with $\mu = 1000(.001) = 1$, which is obtained as follows:

$$Pr(X \geq 4) = 1 - [Pr(X = 0) + Pr(X = 1) + Pr(X = 2) + Pr(X = 3)]$$

Using Table 2 of the Appendix under the $\mu = 1.0$ column, we find that

$$Pr(X = 0) = .3679$$

$$Pr(X = 1) = .3679$$

$$Pr(X = 2) = .1839$$

$$Pr(X = 3) = .0613$$

Thus,

$$\begin{aligned} Pr(X \geq 4) &= 1 - (.3679 + .3679 + .1839 + .0613) \\ &= 1 - .9810 = .0190 \end{aligned}$$

This event is indeed unusual and suggests a genetic susceptibility to breast cancer among female offspring of women who have had breast cancer. ■■■

How large should n be or how small should p be before the approximation is "adequate"? A conservative rule is to use the approximation when $n \geq 100$ and $p \leq .01$. As an example we give the exact binomial probability and the Poisson approximation for $n = 100$, $p = .01$, $k = 0, 1, 2, 3, 4, 5$ in Table 4.6. The two probability distributions agree to within .002 in all instances.

TABLE 4.6 An example of the Poisson approximation to the binomial distribution for $n = 100$, $p = .01$, $k = 0, 1, \ldots, 5$

k	Exact binomial probability	Poisson approximation	k	Exact binomial probability	Poisson approximation
0	.366	.368	3	.061	.061
1	.370	.368	4	.015	.015
2	.185	.184	5	.003	.003

SECTION 4.14 Summary

In this chapter, random variables were discussed and a distinction between discrete and continuous random variables was made. Specific attributes of random variables, including the notions of probability mass function (or probability distribution), cumulative-distribution function, expected value, and variance were introduced. These notions were shown to be related to similar concepts for finite samples, which were discussed in Chapter 2. In particular, the sample frequency distribution is a sample realization of a probability distribution, whereas the sample mean ($\bar{x}$) and variance (s^2) are sample analogues of the expected value and variance, respectively, of a random variable. The relationship between attributes of probability models and finite samples is explored in more detail in Chapter 6.

Finally, some specific probability models were introduced, focusing on the binomial and Poisson distributions. The binomial distribution was shown to be applicable for binary outcomes, that is, if only two outcomes are possible, where outcomes on different trials are independent. These two outcomes are labeled as "success" and "failure," where the probability of success is the same for each trial. The Poisson distribution is a classic model used to describe the distribution of rare events.

The study of probability models continues in Chapter 5, where the focus is on continuous random variables.

PROBLEMS

Let X be the random variable representing the number of hypertensive adults in Example 3.13.

* **4.1** Derive the probability mass function for X.

* **4.2** What is its expected value?

* **4.3** What is its variance?

* **4.4** What is the cumulative-distribution function?

Suppose we wish to check the accuracy of self-reported diagnoses of angina by getting further medical records on a subset of the cases.

4.5 If we have 50 reported cases of angina and we wish to select 5 for further review, then how many ways can we select these cases if the order of selection matters?

4.6 Answer Problem 4.5 if the order of selection does not matter.

4.7 Evaluate ${}_{10}C_0$, ${}_{10}C_1$, . . . , ${}_{10}C_{10}$.

* **4.8** Evaluate 9!.

4.9 Suppose that 6 out of 15 students in a grade-school class develop influenza, whereas 20% of grade-school students nationwide develop influenza. Is there evidence of an excessive number of cases in the class? That is, what is the probability of obtaining at least 6 cases in this class if the nationwide rate holds true?

4.10 What is the expected number of students in the class who will develop influenza?

* **4.11** What is the probability of obtaining exactly 6 events for a Poisson distribution with parameter $\mu = 4.0$?

* **4.12** What is the probability of obtaining at least 6 events for a Poisson distribution with parameter $\mu = 4.0$?

* **4.13** What is the expected value and variance for a Poisson distribution with parameter $\mu = 4.0$?

Infectious Disease

Newborns were screened for human immunodeficiency virus (HIV or AIDS virus) in five Massachusetts hospitals. The data obtained [7] are shown in Table 4.7.

4.14 If 500 newborns are screened at the inner-city hospital, then what is the exact binomial probability of precisely 5 HIV-positive test results?

4.15 If 500 newborns are screened at the inner-city hospital, then what is the exact binomial probability of at least 5 HIV-positive test results?

TABLE 4.7 Seroprevalence of HIV antibody in newborns' blood samples, according to hospital category

Hospital	Type	Number tested	Number positive	Number positive (per 1000)
A	Inner city	3,741	30	8.0
B	Urban/Suburban	11,864	31	2.6
C	Urban/Suburban	5,006	11	2.2
D	Suburban/Rural	3,596	1	0.3
E	Suburban/Rural	6,501	8	1.2

4.16 Answer Problems 4.14 and 4.15 using an approximation rather than an exact probability.

4.17 Answer Problem 4.14 for a mixed urban/suburban hospital (hospital C).

4.18 Answer Problem 4.15 for a mixed urban/suburban hospital (hospital C).

4.19 Answer Problem 4.16 for a mixed urban/suburban hospital (hospital C).

4.20 Answer Problem 4.14 for a mixed suburban/rural hospital (hospital E).

4.21 Answer Problem 4.15 for a mixed suburban/rural hospital (hospital E).

4.22 Answer Problem 4.16 for a mixed suburban/rural hospital (hospital E).

Occupational Health

Many investigators have suspected that workers in the tire industry have an unusual incidence of cancer.

* **4.23** Suppose the expected number of deaths due to bladder cancer for all workers in a tire plant on January 1, 1964, over the next 20 years (1/1/64–12/31/83) based on U.S. mortality rates is 1.8. If the Poisson distribution is assumed to hold and there are 6 reported deaths due to bladder cancer among the tire workers, then how unusual is this event?

* **4.24** Suppose a similar analysis is done for stomach cancer. In this plant, 4 deaths due to stomach cancer are observed for the workers, whereas 2.5 are expected based on U.S. mortality rates. How unusual is this event?

Infectious Disease

One hypothesis is that gonorrhea tends to cluster in central cities.

4.25 Suppose that 10 gonorrhea cases are reported over a 3-month period among 10,000 people living in an urban county. The statewide incidence of gonorrhea is 50 per 100,000 over a 3-month period. Is the number of gonorrhea cases in this county unusual for this time period?

Otolaryngology

Assume that the number of episodes per year of otitis media, a common disease of the middle ear in early childhood, follows a Poisson distribution with parameter $\lambda = 1.6$.

* **4.26** Find the probability of getting 3 or more episodes of otitis media in the first 2 years of life.

* **4.27** Find the probability of not getting any episodes of otitis media in the first year of life.

An interesting question in pediatrics is whether the tendency for children to have many episodes of otitis media is inherited in a family.

* **4.28** What is the probability that 2 siblings will both have 3 or more episodes of otitis media in the first 2 years of life?

* **4.29** What is the probability that exactly 1 of the siblings will have 3 or more episodes in the first 2 years of life?

* **4.30** What is the probability that neither sibling will have 3 or more episodes in the first 2 years of life?

* **4.31** What is the expected number of siblings in a 2-sibling family that will have 3 or more episodes in the first 2 years of life?

Hypertension

Hypertension has often been claimed to have a "familial aggregation." That is, if 1 person in a family is hypertensive, then his or her siblings are more likely to be hypertensive. Suppose that the prevalence of hypertension among 50–59-year-olds in the general population is 18%. Suppose we identify sibships of size 3 in a community where all members of the sibship are 50–59 years old.

4.32 What is the probability that 0, 1, 2, or 3 hypertensives will be identified in such sibships if the hypertensive status of 2 siblings in the same family are independent events?

4.33 Suppose that among 25 sibships of this type, 5 have at least 2 affected siblings. Are these data consistent with the independence assumption in Problem 4.32?

Environmental Health, Obstetrics

Suppose that the rate of major congenital malformations in the general population is 2.5 per 100 deliveries. A study is set up to investigate if the offspring of Vietnam-veteran fathers are at special risk of having congenital malformations.

* **4.34** If 100 infants are identified in a birth registry as being offspring of a Vietnam-veteran father and 4 have a major congenital malformation, then is there an excess risk of malformations in this group?

Using these same birth-registry data, let us look at the effect of maternal use of marijuana on the rate of major congenital malformations.

* **4.35** Of 75 offspring of mothers who used marijuana, 8 are found to have a major congenital malformation. Is there an excess risk of malformations in this group?

Hypertension

A national study found that treating people appropriately for high blood pressure reduced their overall mortality by 20%. Treating people adequately for hypertension has been difficult, since it is estimated that 50% of hypertensives do not know they have high blood pressure; 50% of those that do know are inadequately treated by their physicians; and 50% that are appropriately treated fail to comply with this treatment by taking the appropriate number of pills.

4.36 What is the probability that among 10 true hypertensives at least 50% are being treated appropriately and are complying with this treatment?

4.37 What is the probability that at least 7 of the 10 hypertensives know they have high blood pressure?

4.38 If the preceding 50% rates were each reduced to 40% by a massive education program, then what effect would this rate change have on the overall mortality rate among true hypertensives; that is, would the mortality rate decrease, and if so, by what percent?

Renal Disease

The presence of bacteria in a urine sample (bacteriuria) is sometimes associated with symptoms of kidney disease in women. Suppose that a determination of bacteriuria has been made over a large population of women at one point in time and that 5% of those sampled are positive for bacteriuria.

* **4.39** If a sample of size 5 is selected from this population, what would be the probability that 1 or more women would be positive for bacteriuria?

* **4.40** Suppose 100 women from this population are sampled. What is the probability that 3 or more women would be positive for bacteriuria?

One interesting phenomenon of bacteriuria is that there is a "turnover"; that is, if bacteriuria is measured on the same woman at 2 different points in time, the results are not necessarily the same. Assume that 20% of all women who are bacteriuric at time 0 are again bacteriuric at time 1 (1 year later), whereas only 4.2% of women who were not bacteriuric at time 0 *are* bacteriuric at time 1. Let X be the random variable representing the number of bacteriuric events over the 2 time periods for 1 woman and still assume that the probability that a woman will be positive for bacteriuria at any one exam is 5%.

* **4.41** What is the probability distribution of X?

* **4.42** What is the mean of X?

* **4.43** What is the variance of X?

Demography

In Table 4.8 we provide life-table data for the United States in 1986 [3]. This table can be used to estimate the probability of survival between any two ages for persons of a given race or sex. For example, for white males, to calculate the probability of survival from age 60 to age 62, we refer to the age 60 and 62 lines under the white male column and obtain a probability of 79,669/82,435 = .966. Refer to the 11 males among the 25 people described in Table 2.11. (The race of the subjects is not known, so use the "All races" section of Table 4.8.)

4.44 What is the expected number of deaths among the 11 males over the next year based on the life-table data?

4.45 Answer Problem 4.44 for a 2-year period.

4.46 Answer Problem 4.44 for a 3-year period.

Use a computer, if necessary, to answer Problems 4.47–4.52.

4.47 What is the probability of exactly 2 deaths among the 11 males over the next year?

4.48 Answer Problem 4.47 for a 2-year period.

4.49 Answer Problem 4.47 for a 3-year period.

TABLE 4.8 Number of survivors at single years of age, out of 100,000 born alive, by race and sex: United States, 1986

Age	All races			White			All other					
							Total			Black		
	Both sexes	Male	Female	Both sexes	Male	Female	Both sexes	Male	Female	Both sexes	Male	Female
0	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000
1	98,964	98,845	99,090	99,106	98,998	99,220	98,426	98,262	98,597	98,190	97,996	98,391
2	98,892	98,764	99,028	99,040	98,923	99,164	98,332	98,158	98,513	98,085	97,882	98,296
3	98,838	98,704	98,980	98,991	98,869	99,121	98,256	98,075	98,444	98,000	97,790	98,218
4	98,796	98,658	98,942	98,953	98,828	99,087	98,194	98,007	98,388	97,932	97,716	98,155
5	98,762	98,620	98,912	98,923	98,794	99,060	98,143	97,951	98,342	97,876	97,655	98,104
6	98,733	98,587	98,887	98,897	98,765	99,038	98,100	97,904	98,304	97,830	97,605	98,062
7	98,707	98,557	98,866	98,874	98,738	99,019	98,064	97,863	98,272	97,792	97,563	98,028
8	98,684	98,530	98,847	98,853	98,712	99,002	98,033	97,828	98,246	97,760	97,526	97,999
9	98,663	98,505	98,830	98,834	98,689	98,987	98,006	97,797	98,223	97,732	97,494	97,975
10	98,645	98,484	98,815	98,817	98,669	98,973	97,982	97,769	98,203	97,706	97,464	97,954
11	98,628	98,464	98,801	98,802	98,651	98,960	97,959	97,742	98,184	97,681	97,435	97,934
12	98,610	98,443	98,786	98,785	98,632	98,946	97,935	97,713	98,166	97,655	97,404	97,914
13	98,587	98,415	98,768	98,763	98,606	98,929	97,907	97,677	98,146	97,625	97,365	97,893
14	98,554	98,372	98,745	98,730	98,564	98,906	97,871	97,629	98,123	97,587	97,314	97,869
15	98,507	98,309	98,716	98,683	98,501	98,876	97,825	97,565	98,095	97,538	97,247	97,840
16	98,445	98,223	98,678	98,620	98,414	98,838	97,767	97,484	98,061	97,477	97,161	97,805
17	98,369	98,116	98,633	98,542	98,306	98,792	97,696	97,384	98,021	97,403	97,056	97,763
18	98,280	97,991	98,583	98,452	98,180	98,741	97,612	97,264	97,975	97,315	96,930	97,715
19	98,183	97,852	98,530	98,355	98,041	98,688	97,515	97,123	97,923	97,214	96,781	97,662
20	98,081	97,702	98,477	98,254	97,894	98,635	97,405	96,959	97,867	97,098	96,607	97,603
21	97,974	97,542	98,424	98,150	97,740	98,583	97,281	96,771	97,805	96,967	96,407	97,538
22	97,861	97,372	98,370	98,043	97,578	98,532	97,143	96,560	97,737	96,821	96,181	97,467
23	97,745	97,196	98,316	97,935	97,412	98,482	96,993	96,331	97,664	96,661	95,933	97,390
24	97,628	97,018	98,261	97,827	97,247	98,432	96,834	96,089	97,586	96,490	95,670	97,306
25	97,512	96,843	98,204	97,720	97,086	98,381	96,669	95,839	97,502	96,311	95,395	97,216
26	97,397	96,672	98,147	97,616	96,931	98,330	96,499	95,582	97,413	96,124	95,110	97,119
27	97,283	96,504	98,088	97,514	96,780	98,278	96,322	95,317	97,318	95,927	94,814	97,015
28	97,168	96,336	98,027	97,412	96,632	98,225	96,137	95,043	97,217	95,719	94,503	96,902
29	97,050	96,164	97,964	97,309	96,481	98,171	95,942	94,756	97,107	95,497	94,174	96,778
30	96,927	95,986	97,897	97,202	96,325	98,115	95,735	94,455	96,988	95,258	93,823	96,642
31	96,798	95,800	97,826	97,090	96,162	98,056	95,515	94,139	96,858	95,001	93,450	96,492
32	96,663	95,606	97,751	96,973	95,993	97,994	95,283	93,807	96,718	94,727	93,054	96,329
33	96,522	95,405	97,672	96,852	95,818	97,928	95,038	93,459	96,568	94,436	92,636	96,154
34	96,377	95,199	97,588	96,727	95,639	97,859	94,781	93,093	96,411	94,131	92,196	95,970
35	96,227	94,988	97,500	96,598	95,457	97,786	94,512	92,708	96,247	93,812	91,735	95,779
36	96,072	94,771	97,407	96,465	95,271	97,708	94,230	92,303	96,076	93,480	91,252	95,581
37	95,910	94,546	97,308	96,326	95,079	97,625	93,934	91,877	95,896	93,132	90,745	95,373
38	95,739	94,311	97,202	96,180	94,878	97,534	93,622	91,429	95,705	92,765	90,212	95,152
39	95,557	94,065	97,086	96,024	94,667	97,434	93,291	90,958	95,498	92,373	89,649	94,911
40	95,363	93,805	96,958	95,856	94,443	97,323	92,938	90,462	95,271	91,952	89,052	94,646
41	95,153	93,528	96,816	95,674	94,203	97,199	92,560	89,939	95,022	91,500	88,419	94,353
42	94,927	93,232	96,659	95,476	93,945	97,061	92,156	89,386	94,751	91,015	87,749	94,033
43	94,683	92,916	96,487	95,261	93,668	96,909	91,727	88,806	94,458	90,501	87,046	93,687
44	94,421	92,579	96,300	95,029	93,371	96,743	91,276	88,201	94,145	89,962	86,315	93,318
45	94,139	92,218	96,097	94,778	93,051	96,563	90,803	87,572	93,812	89,400	85,559	92,928
46	93,836	91,831	95,878	94,507	92,705	96,367	90,307	86,919	93,458	88,816	84,781	92,518
47	93,508	91,413	95,639	94,212	92,329	96,153	89,784	86,237	93,081	88,206	83,974	92,082
48	93,151	90,960	95,377	93,888	91,919	95,917	89,226	85,514	92,673	87,559	83,125	91,614
49	92,758	90,464	95,087	93,531	91,469	95,656	88,620	84,733	92,226	86,862	82,217	91,104
50	92,325	89,919	94,766	93,137	90,973	95,365	87,957	83,881	91,733	86,104	81,237	90,545
51	91,848	89,320	94,409	92,701	90,427	95,042	87,230	82,951	91,189	85,280	80,178	89,931
52	91,324	88,665	94,016	92,221	89,827	94,684	86,440	81,945	90,594	84,389	79,042	89,261
53	90,752	87,949	93,586	91,694	89,167	94,291	85,591	80,869	89,951	83,437	77,836	88,540
54	90,130	87,169	93,122	91,117	88,441	93,864	84,693	79,735	89,268	82,433	76,575	87,774
55	89,458	86,322	92,623	90,488	87,644	93,401	83,749	78,549	88,548	81,382	75,266	86,968
56	88,733	85,405	92,088	89,803	86,773	92,901	82,763	77,316	87,794	80,287	73,915	86,124
57	87,951	84,413	91,513	89,058	85,823	92,360	81,729	76,030	87,000	79,142	72,514	85,236
58	87,103	83,339	90,888	88,247	84,788	91,770	80,630	74,674	86,148	77,930	71,046	84,284
59	86,180	82,173	90,203	87,361	83,661	91,121	79,444	73,227	85,215	76,626	69,487	83,244
60	85,173	80,908	89,449	86,393	82,435	90,406	78,156	71,675	84,185	75,215	67,822	82,097
61	84,077	79,539	88,619	85,338	81,105	89,619	76,757	70,011	83,047	73,687	66,042	80,834
62	82,891	78,065	87,712	84,193	79,669	88,758	75,254	68,243	81,808	72,051	64,157	79,461
63	81,618	76,492	86,731	82,961	78,131	87,823	73,665	66,388	80,485	70,328	62,189	77,999
64	80,264	74,827	85,681	81,645	76,497	86,817	72,016	64,473	79,104	68,548	60,167	76,479
65	78,833	73,076	84,565	80,246	74,770	85,740	70,325	62,516	77,684	66,732	58,113	74,922
66	77,327	71,244	83,381	78,766	72,955	84,590	68,600	60,526	76,232	64,890	56,039	73,337
67	75,740	69,325	82,122	77,199	71,046	83,360	66,834	58,499	74,737	63,015	53,941	71,715
68	74,059	67,305	80,776	75,532	69,027	82,040	65,011	56,423	73,180	61,092	51,805	70,035
69	72,267	65,161	79,330	73,750	66,877	80,618	63,109	54,279	71,534	59,098	49,612	68,270
70	70,353	62,881	77,772	71,841	64,581	79,082	61,111	52,055	69,778	57,018	47,350	66,401
71	68,312	60,462	76,096	69,801	62,139	77,427	59,013	49,749	67,903	54,848	45,018	64,419
72	66,149	57,916	74,299	67,634	59,561	75,649	56,823	47,373	65,916	52,598	42,630	62,332
73	63,872	55,256	72,383	65,347	56,860	73,748	54,557	44,945	63,830	50,284	40,203	60,154
74	61,491	52,503	70,350	62,949	54,057	71,724	52,236	42,489	61,666	47,926	37,763	57,904
75	59,016	49,675	68,200	60,450	51,169	69,577	49,877	40,023	59,437	45,540	35,329	55,597
76	56,452	46,785	65,931	57,853	48,210	67,303	47,486	37,558	57,145	43,133	32,914	53,236
77	53,800	43,842	63,538	55,162	45,191	64,896	45,061	35,097	54,783	40,706	30,522	50,815
78	51,061	40,855	61,012	52,376	42,122	62,350	42,594	32,639	52,339	38,254	28,154	48,324
79	48,235	37,833	58,347	49,498	39,013	59,657	40,078	30,181	49,799	35,773	25,811	45,749
80	45,324	34,789	55,535	46,530	35,879	56,812	37,506	27,723	47,151	33,260	23,495	43,081
81	42,333	31,739	52,570	43,478	32,737	53,809	34,876	25,270	44,387	30,716	21,215	40,314
82	39,269	28,705	49,450	40,350	29,611	50,643	32,192	22,831	41,506	28,147	18,982	37,447
83	36,144	25,712	46,172	37,158	26,528	47,313	29,461	20,419	38,509	25,564	16,812	34,485
84	32,972	22,791	42,736	33,916	23,520	43,817	26,695	18,053	35,405	22,982	14,728	31,435
85	29,771	19,977	39,143	30,642	20,625	40,155	23,912	15,755	32,206	20,419	12,755	28,312

4.50 What is the probability of at least 4 deaths among the 11 males over the next year?

4.51 Answer Problem 4.50 for a 2-year period.

4.52 Answer Problem 4.50 for a 3-year period.

Pediatrics, Otolaryngology

Otitis media is a disease that occurs frequently in the first few years of life and is one of the most common reasons for physician visits after the routine check-up. A study was conducted to assess the frequency of otitis media in the general population in the first year of life. Table 4.9 gives the number of infants out of 2500 infants who were first seen at birth and who remained disease-free by the end of the ith month of life, $i = 0, 1, \ldots, 12$. (Assume that no infants have been lost to follow-up.)

* **4.53** What is the probability that an infant will have 1 or more episodes of otitis media by the end of the 6th month of life? the first year of life?

TABLE 4.9 Number of infants (out of 2500) who remain disease-free at the end of each month during the first year of life

i	Disease-free infants at the end of month i
0	2500
1	2425
2	2375
3	2300
4	2180
5	2000
6	1875
7	1700
8	1500
9	1300
10	1250
11	1225
12	1200

* **4.54** What is the probability that an infant will have 1 or more episodes of otitis media by the end of the 9th month of life given that no episodes have been observed by the end of the 3rd month of life?

* **4.55** Suppose an "otitis-prone family" is defined as one where at least 3 siblings out of 5 develop otitis media in the first 6 months of life. What proportion of 5-sibling families are otitis prone if we assume that the disease occurs independently for different siblings in a family?

* **4.56** What is the expected number of otitis-prone families out of 100 5-sibling families?

Cancer, Epidemiology

An experiment is designed to test the potency of a drug on 20 rats. Previous animal studies have shown that a 10-mg dose of the drug is lethal 5% of the time within the first 4 hours; of the animals alive at 4 hours, 10% will die in the next 4 hours.

4.57 What is the probability that 3 or more rats will die in the first 4 hours?

4.58 Suppose 2 rats die in the first 4 hours. What is the probability that 2 or fewer rats will die in the next 4 hours?

4.59 What is the probability that 0 rats will die in the 8-hour period?

4.60 What is the probability that 1 rat will die in the 8-hour period?

4.61 What is the probability that 2 rats will die in the 8-hour period?

4.62 Can you write a general formula for the probability that x rats will die in the 8-hour period? Evaluate this formula for $x = 0, 1, \ldots, 10$.

Environmental Health

One of the important issues in assessing nuclear energy is whether there are excess disease risks in the communities surrounding nuclear-power plants. A study was undertaken in the community surrounding Hanford, Washington, looking at the prevalence of selected congenital malformations in the counties surrounding the nuclear-test facility [8].

* **4.63** Suppose that 27 cases of Downs syndrome are found and only 19 are expected based on Birth Defects Monitoring Program prevalence estimates conducted in the states of Washington, Idaho, and Oregon. Is there a significant excess number of cases in the area surrounding the nuclear-power plant?

Suppose that 12 cases of cleft palate are observed, while only 7 are expected based on Birth Defects Monitoring Program estimates.

* **4.64** What is the probability of observing exactly 12 cases of cleft palate if there is no excess risk of cleft palate in the study area?

* **4.65** Do you feel there is a meaningful excess number of cases of cleft palate in the area surrounding the nuclear-power plant?

Health Promotion

A study was conducted among 234 people who had expressed a desire to stop smoking but who had not yet stopped. On the day they quit smoking, their carbon-monoxide level (CO) was measured and the time was noted from the time they smoked their last cigarette to the time of the CO measurement. The CO level provides an "objective" indicator of the number of cigarettes smoked per day during the time immediately prior to the quit attempt. However, it is known to also be influenced by the time since the last cigarette was smoked. Thus, this time is provided as well as a "corrected CO level," which is adjusted for the time since last smoked. Information is also provided on the age and sex of the subjects as well as the subject's self-report of the number of cigarettes per day. The subjects were followed up for one year for the purpose of determining the number of days they remained abstinent. The number of days abstinent ranges from 0 days for those who quit for less than 1 day to 365 days for those who were abstinent for the full year. Assume that all persons were followed for the entire year.

The data are given in Data Set SMOKE.DAT, on the data disk. The format of this file is given in Table 4.10.

4.66 Develop a life table similar to Table 4.9, giving the number of persons who remained abstinent at 1, 2, . . . , 12 months of life (assume for simplicity that there are 30 days in each of the first 11 months after quitting and 35 days in the 12th month). Plot these data either by hand or on the computer. Compute the probability that a person will remain abstinent at 1, 3, 6, and 12 months after quitting.

4.67 Develop life tables for subsets of the data based on age, sex, number of cigarettes per day, and carbon-monoxide level (one variable at a time). Based on these data, do you feel that age, sex, number of cigarettes per day, or CO level are related to success in quitting? (Methods of analysis for life-table data are discussed in more detail in Chapter 13.)

Genetics

4.68 A topic of some interest in the genetic literature over at least the last 30 years has been the study of sex-ratio data. In particular, one hypothesis that has been suggested is that there are a sufficient number of families with a preponderance of males (females) that the sexes of successive childbirths are not independent random variables but are related to each other. This hypothesis has been extended beyond just successive births so that some authors also consider relationships between offspring two birth orders apart (i.e., 1st and 3rd offspring, 2nd and 4th offspring, etc.). Sex-ratio data from the first 5 births in 51,868 families are given in Data Set SEXRAT.DAT (on the data disk). The format of this file is given in Table 4.11 [9]. What are your conclusions concerning the above hypothesis based on your analysis of these data?

TABLE 4.10 Format of SMOKE.DAT

Variable	Columns	Code
ID number	1–3	
Age	4–5	
Gender	6	1 = male, 2 = female
Cigarettes/day	7–8	
Carbon monoxide (CO) (× 10)	9–11	
Minutes elapsed since the last cigarette smoked	12–15	
LogCOAdj[a] (× 1000)	16–19	
Days abstinent[b]	20–22	

[a] This variable represents adjusted carbon monoxide (CO) values. CO values were adjusted for minutes elapsed since the last cigarette smoked using the formula, $\text{Log}_{10}\text{CO (adjusted)} = \text{Log}_{10}\text{CO} - (-0.000638) \times (\text{min} - 80)$, where min is the number of minutes elapsed since the last cigarette smoked.

[b] Those abstinent less than 1 day were given a value of 0.

TABLE 4.11 Format of SEXRAT.DAT

Variable	Column
Number of children[a]	1
Sex of children[b]	3–7
Number of families	9–12

[a] For families with 5+ children, the sexes of the first 5 children are listed. The number of children is given as 5 for such families.
[b] The sex of successive births is given. Thus, MMMF means that the first 3 children were males and the 4th child was a female. There were 484 such families.

Infectious Disease

A study was conducted of risk factors for HIV infection among intravenous drug users [10]. It was found that 40% of users who had ≤ 100 injections per month (light users) and 55% of users who had > 100 injections per month (heavy users) were HIV positive.

4.69 What is the probability that exactly 3 of 5 light users were HIV positive?

4.70 What is the probability that at least 3 of 5 light users were HIV positive?

4.71 Suppose we have a group of 10 light users and 10 heavy users. What is the probability that exactly 3 of the 20 users will be HIV positive?

4.72 What is the probability that at least 4 of the 20 users will be HIV positive?

Ophthalmology, Diabetes

In a recent study [11] of incidence rates of blindness among insulin-dependent diabetics, it was reported that the annual incidence rate of blindness per year was 0.67% among 30–39-year-old male insulin-dependent diabetics (male IDDM) and 0.74% among 30–39-year-old female insulin-dependent diabetics (female IDDM).

4.73 If a group of 200 IDDM 30–39-year-old men is gathered, what is the probability that exactly 2 will become blind over a 1-year period?

4.74 If a group of 200 IDDM 30–39-year-old women is gathered, what is the probability that at least 2 will become blind over a 1-year period?

4.75 What is the probability that a 30-year-old male IDDM patient will become blind over the next 10 years?

4.76 After how many years of follow-up would we expect the cumulative incidence of blindness to be 10% among 30-year-old IDDM women if the incidence rate remains constant over time?

4.77 What does cumulative incidence mean, in words, in the context of this problem?

References

[1] *Boston Globe*, October 7, 1980.

[2] Rinsky, R. A., Zumwalde, R. O., Waxweiler, R. J., Murray, W. E., Bierbaum, P. J., Landrigan, P. J., Terpilak, M., & Cox, C. (1981, January 31). Cancer mortality at a naval nuclear shipyard. *Lancet*, 231–235.

[3] U.S. Department of Health and Human Services (1986). *Vital Statistics of the United States, 1986*.

[4] National Center for Health Statistics. (1974, June 27). *Monthly vital statistics report, annual summary for the United States (1973)*, *22*(13).

[5] National Center for Health Statistics. (1978, December 7). *Monthly vital statistics report, annual summary for the United States (1977)*, *26*(13).

[6] Ott, M. G., Scharnweber, H. C., & Langner, R. (1980). Mortality experience of 161 employees exposed to ethylene dibromide in two production units. *British Journal of Industrial Medicine*, *37*, 163–168.

[7] Hoff, R., Berardi, V. P., Weiblen, B. J., Mahoney-Trout, L., Mitchell, M. L., & Grady, G. F. (1988). Seroprevalence of human immunodeficiency virus among childbearing women. *New England Journal of Medicine*, *318*(9), 525–530.

[8] Sever, L. E., Hessol, N. A., Gilbert, E. S., & McIntyre, J. M. (1988). The prevalence at birth of congenital malformations in communities near the Hanford site. *American Journal of Epidemiology*, *127*(2), 243–254.

[9] Renkonen, K. O., Mäkelä, O., & Lehtovaara, R. (1961). Factors affecting the human sex ratio. *Annales Medicinae Experimentalis et Biologiae Fenniae*, *39*, 173–184.

[10] Schoenbaum, E. E., Hartel, D., Selwyn, P. A., Klein, R. S., Davenny, K., Rogers, M., Feiner, C., & Friedland, G. (1989). Risk factors for human immunodeficiency virus infection in intravenous drug users. *New England Journal of Medicine*, *321*(13), 874–879.

[11] Sjolie, A. K., & Green, A. (1987). Blindness in insulin-treated diabetic patients with age at onset less than 30 years. *Journal of Chronic Disease*, *40*(3), 215–220.