Chapter 4

Discrete Probability Distributions

4.1 Introduction

In this chapter, we will discuss

- Discrete random variables
- Probability-mass function (probability distribution)
- Binomial distribution

4.2 Random Variables

A random variable is function that assigns numeric values to different events in a sample space.

Two types of random variables: discrete and continuous

- A random variable for which there exists a discrete set of numeric values is a discrete random variable.

- A random variable whose possible values cannot be enumerated is a continuous random variable.

EXAMPLE 4.3

Otolaryngology Otitis media, a disease of the middle ear, is one of the most common 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, and so on.

EXAMPLE 4.4

Hypertension Many new drugs have been introduced in the past several decades 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 four untreated hypertensives she encounters in her practice, before deciding whether to adopt the drug for routine use. Let X = the number of patients of four who are brought under control. Then X is a discrete random variable, which takes on the values 0, 1, 2, 3, 4.

EXAMPLE 4.5

Environmental Health Possible health effects on workers of exposure to low levels of radiation over long periods of time are of public health interest. One problem in assessing this issue is how to measure the cumulative exposure of a worker. A study was performed at the Portsmouth Naval Shipyard, where 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. Cumulative lifetime exposure to radiation is a good example of a continuous random variable because it varied in this study from 0.000 to 91.414 rem; this would be regarded as taking on an essentially infinite number of values, which cannot be enumerated.

4.3 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 called a **probability-mass function** (pmf).

A **probability-mass function**, sometimes also called a **probability distribution**, 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.

EXAMPLE 4.6 **Hypertension** Consider Example 4.4. Suppose from previous experience with the drug, the drug company expects that for any clinical practice the probability that 0 patients of 4 will be brought under control is .008, 1 patient of 4 is .076, 2 patients of 4 is .265, 3 patients 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, $0 < Pr(X = r) \leq 1$, $\sum Pr(X = r) = 1$, where the summation is taken over all possible values that have positive probability.

Relationship of Probability Distributions to Frequency Distributions

Frequency distribution is described as a list of each value in the data set and a corresponding count of how frequently the value occurs.

If each count is divided by the total number of points in the sample, then the frequency distribution can be considered as a sample analog to a probability distribution.

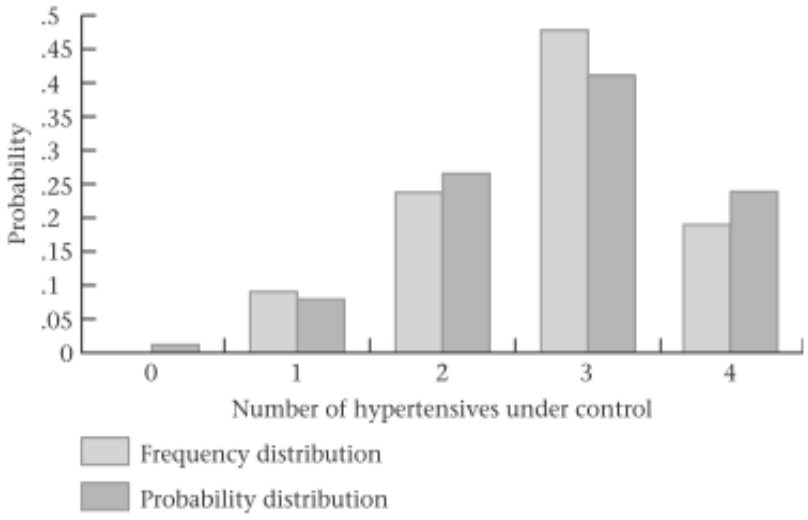
Frequency distribution gives the actual proportion of points in a sample that correspond to specific values. Hence, the appropriateness of a model can be assessed by comparing the observed sample-frequency distribution with the probability distribution, also called as **goodness-of-fit**.

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

Pmf derived from the binomial distribution is compared with the frequency distribution to determine whether the drug behaves with the same efficacy as predicted.

Figure 4.1 Comparison of the frequency and probability distribution for the hypertension-control example



4.4Expected Value of a Discrete Random Variable

- If a random variable has a large number of values with positive probability then the pmf is not a useful measure.
- Measures of location and spread can be developed for a random variable in much the same way as for samples.
- The analog of the arithmetic mean $\bar{x}$ is called the expected value of a random variable, or population mean, and is denoted by E(X) or μ and represents the "average" value of the random variable.
- The expected value of a discrete random variable is defined as

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

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

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 average a child would be expected to have about two episodes of otitis media in the first 2 years of life.

4.5 Variance of a Discrete Random Variable

The analog of the sample variance (s^2) for a random variable is called the **variance of a random variable, or population variance**, and is denoted by Var(X) or σ^2

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

OR

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

where x_i^2 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.

The variance represents the spread, relative to the expected value, of all values that have positive probability.

Approximately 95% of the probability mass falls within two standard deviations (2σ) of the mean of a random variable.

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 $\mu = 2.038$. Furthermore,

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

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

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 (μ) = 2.038 and standard deviation (σ) = 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. Because only positive-integer values are possible for this random variable, the valid range is from $a = 0$ to $b = 4$ episodes. Table 4.3 gives the probability of having ≤4 episodes as

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

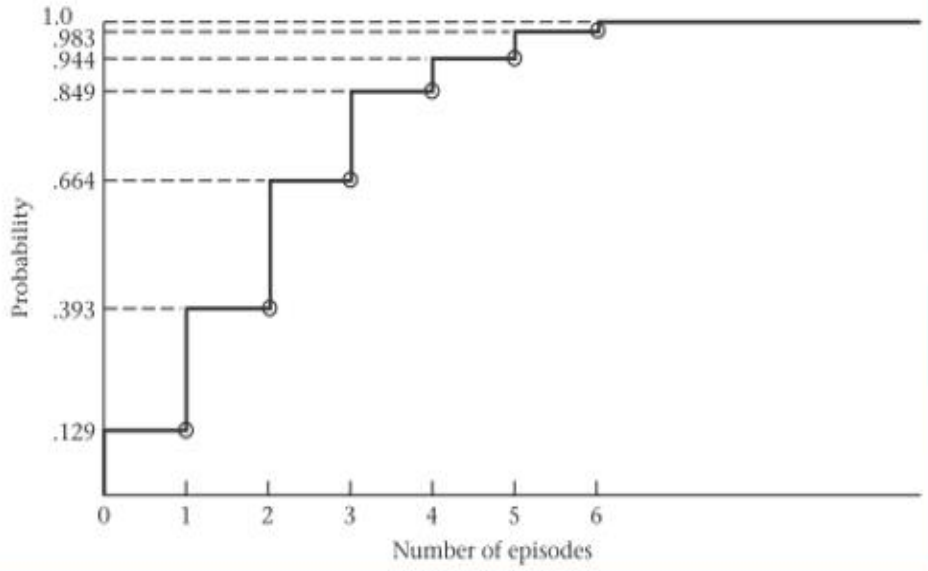
The rule lets us quickly summarize the range of values that have most of the probability mass for a random variable without specifying each individual value. Chapter 6 discusses the type of random variable to which Equation 4.2 applies.

Cumulative-Distribution Function of a Discrete Random Variable

The cumulative-distribution function (cdf) of a random variable X is denoted by F(X) and, for a specific value of x of X, is defined by $Pr(X \leq x)$ and denoted by $F(x)$.

Discrete and continuous random variables can be distinguished based on each variable's cdf.

Figure 4.2 Cumulative-distribution function for the number of episodes of otitis media in the first 2 years of life



- For a discrete random variable, the cdf looks like a series of steps, called the step function.

With the increase in number of values, the cdf approaches that of a smooth curve.

- For a continuous random variable, the cdf is a smooth curve.

4.8 Binomial Distribution

A sample of n independent trials, each of which can have only two possible outcomes, which are denoted as “success” and “failure.”

The probability of a success at each trial is assumed to be some constant p, and hence the probability failure at each trial is

1 – p = q.

That is, the probability of k successes within n trials

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

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 of a blood count are (1) counting the number of white blood cells (the "white count") and (2) differentiating the white blood cells that do exist into five categories—namely, neutrophils, lymphocytes, monocytes, eosinophils, and basophils (called the "differential"). Both the white count and the differential are used extensively in making clinical diagnoses. We 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.

EXAMPLE 4.25

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 non-neutrophils, given a probability of .6 that any one cell is a neutrophil?

Solution: If a neutrophil is denoted by an x and a non-neutrophil by an o, then the question being asked is: What is the probability of the outcome $oxoox = Pr(oxoox)$? Because 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 4.26 **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 10 possible orderings that result in 2 neutrophils. Table 4.5 gives the 10 possible orderings.

TABLE 4.5 **Possible orderings for 2 neutrophils 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 $= \binom{5}{2} = (5\times 4)/(2\times 1) = 10.$

The probability of any of the orderings in Table 4.5 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 $\binom{5}{2}(.6)^2(.4)^3 = 10(.6)^2(.4)^3 = .230.$

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

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

Let X be a binomial random variable with parameters n and p.

Let Y be a binomial random variable with parameters n and q = 1 – p.

Then,

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

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*.

EXAMPLE 4.27

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 genders of successive children are considered independent random variables?

Solution: Use a binomial distribution with $n = 5$, $p = .51$, $k = 2$. Let X = number of boys out of 5 births. Compute

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

$$= 10(.51)^2(.49)^3 = .306$$

Using Binomial Tables

Using Binomial Tables

Often a number of binomial probabilities need to be evaluated for the same n and p, which would be tedious if each probability had to be calculated from Equation 4.5. Instead, for small n ($n \leq 20$) and selected values of p, refer to Table 1 in the Appendix, where 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 4.28

Infectious Disease Evaluate the probability of 2 lymphocytes out of 10 white blood cells if the probability of any one cell being 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 4.30

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

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

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

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

$Pr(X = 1) = Pr(Y = 4) = .0768$ on referring to the 4 row and .40 column under $n = 5$
$Pr(X = 2) = Pr(Y = 3) = .2304$ on referring to the 3 row and .40 column under $n = 5$
$Pr(X = 3) = Pr(Y = 2) = .3456$ on referring to the 2 row and .40 column under $n = 5$
$Pr(X = 4) = Pr(Y = 1) = .2592$ on referring to the 1 row and .40 column under $n = 5$
$Pr(X = 5) = Pr(Y = 0) = .0778$ on referring to the 0 row and .40 column under $n = 5$

Using Electronic Tables

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.
If the sample size is not large enough to use normal approximation, then an electronic table can be used to evaluate binomial probabilities.

MS Excel provides a menu of statistical function, including calculation of probabilities for many probability distributions. Example, the binomial-distribution function, called BINOMDIST, which can be used to calculate the pmf and cdf for any binomial distribution.

4.9 Expected Value and Variance of the Binomial Distribution

The expected value of a discrete random variable is

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

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

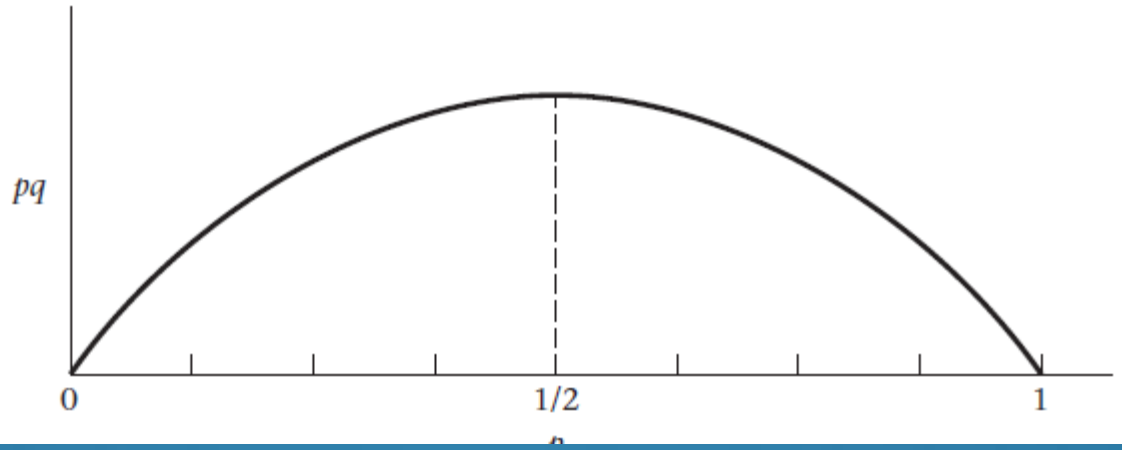
$$\binom{n}{0} p^0 q^n, \quad \binom{n}{1} p^1 q^{n-1}, \ldots$$

Thus, $$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}$$ that is, the expected value = $\boldsymbol{E(x) = np}$

Similarly, variance $\boldsymbol{Var(x) = npq}$

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

FIGURE 4.4 Plot of *pq* versus *p*

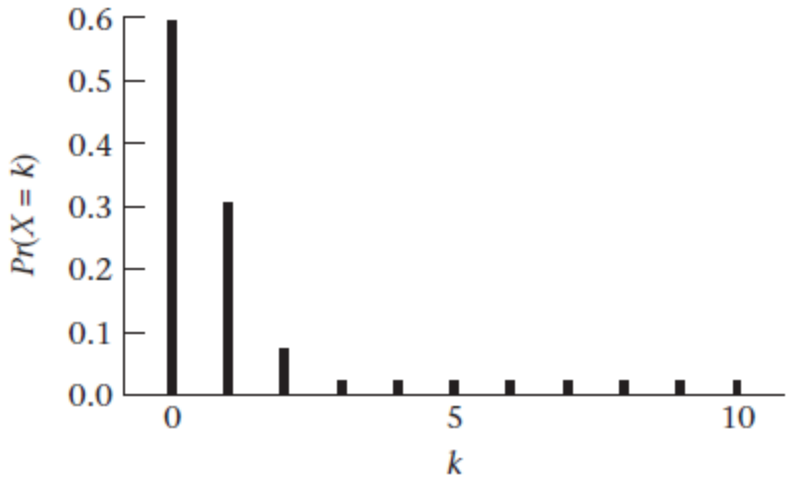


- The expected number of successes in *n* trials is the probability of success in one trial multiplied by n, which equals *np*.

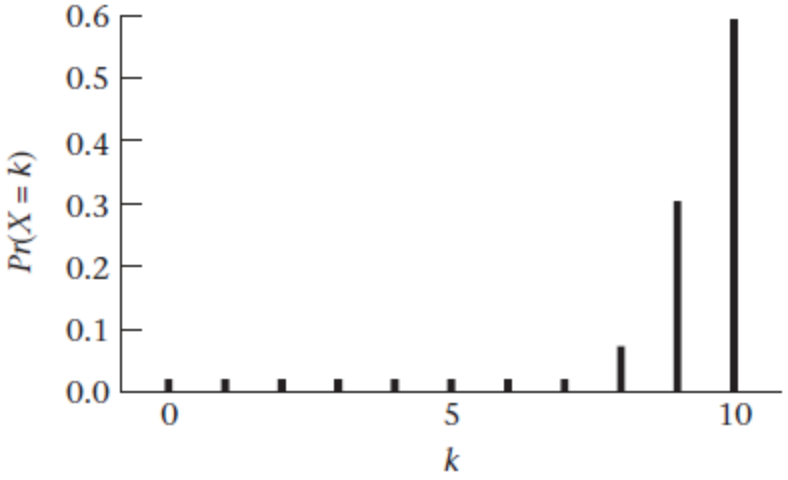
- For a given number of trials *n*, the binomial distribution has the highest variance when p = ½.

- Variance decreases as *p* moves away from ½, becoming 0 when *p* = 0 or 1.

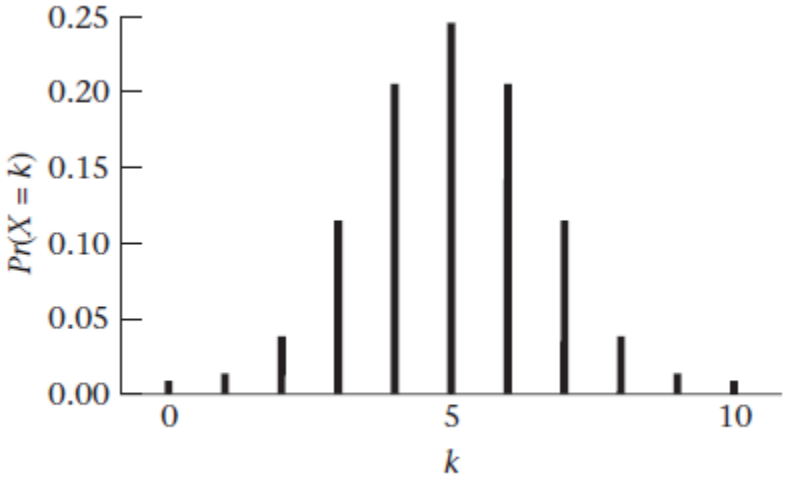
FIGURE 4.5 The binomial distribution for various values of p when $n = 10$



(a) $n = 10$, $p = .05$



(b) $n = 10$, $p = .95$



(c) $n = 10$, $p = .50$

1 The probability of a woman developing breast cancer over a lifetime is about 0.1

(a) What is the probability that exactly 2 women of 10 will develop breast cancer over a lifetime?

(b) What is the probability that at least 2 women of 10 will develop breast cancer over a lifetime?

a) According to what given in the question , the probability of a woman developing breast cancer sometime in her life is about 10% .This means that the probability of a woman not developing breast cancer over a lifetime is about 90%. To calculate the probability that exactly 2 women of 10 will develop breast cancer over a lifetime, we can use the binomial probability formula:

$$P(X = 2) = \binom{10}{2}(0.1)^2(.9)^8$$

where X is the number of women who develop breast cancer, and 0.1 and 0.9 are the probabilities of developing and not developing breast cancer, respectively. Using a calculator, we get: $P(X = 2) = \binom{10}{2}(0.1)^2(.9)^8 = 0.1937$

Therefore, the probability that exactly 2 women of 10 will develop breast cancer over a lifetime is about 0.1937 or 19.37 %. ** Also you can use table 1.**

TABLE 1 Exact binomial probabilities $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$ *(continued)*

n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0313
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051

b) To calculate the probability that at least 2 women of 10 will develop breast cancer over a lifetime, we can use the binomial distribution formula

$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1)$ from table 1

$$1 - 0.3487 - 0.3874 = 0.2639$$

Therefore, the probability that at least 2 women of 10 will develop breast cancer over a lifetime is about 0.2639 or 26.39 %.

Summary

In this chapter, we discussed:

- Random variables and the distinction between discrete and continuous variables.
- Specific attributes of random variables, including notions of probability-mass function (probability distribution), expected value, and variance.
- Sample frequency distribution was described as a sample realization of a probability distribution, whereas sample mean ($\overline{x}$) and variance (s^2) are sample analogs of the expected value and variance, respectively, of a random variable.
- Binomial distribution was shown to be applicable to binary outcomes ("success" and "failure").

The End