13 Random Variables I

> *Dr. Frink*: Mm-hey!! Why, this is the random quote generator I invented! One of my first patents ... aw, look at the garbage they've got you saying!
>
> From: *The Simpsons*

13.1 Introduction

Random variables can be thought of as models of data. They provide a theoretical description of the random processes that might account for the data and help us quantify predictions or decisions we may wish to make based on the data. In fact, many of the tools that we introduced in describing data sets have analogs for random variables. Thus we will discuss histograms, means, and standard deviations for random variables, and these are related to the corresponding concepts for data sets. Some of the tools that we used for representing data, such as medians and quartiles can also be described for random variables, though we will not do so in this course.

In studying data sets we made a broad distinction between univariate data and bivariate or, more generally, multivariate data. Similar distinctions exist for random variables. For example, if we toss a single die we generate an integer X between 1 and 6. This is an example of a univariate random variable, since each observation generates a single number. If our experiment had been to toss a pair of dice (say a red and a white one) and record the outcome, the result would have been a value of a bivariate random variable. In this case we observe a pair of values (X,Y), where X is an integer between 1 and 6 and Y is also an integer between 1 and 6, not necessarily the same as X. Bivariate random variables can be used to model bivariate data. They provide the theoretical underpinnings for making predictions based on correlation and regression analysis. However, that is beyond the scope of these notes.

Besides the distinction between univariate and bivariate random variables, there is an additional classification of random variables into *discrete* and *continuous*. The example of dice tossing in the previous paragraph illustrates the discrete sort. This simply means that we can make a list of all possible values of the random variable. The values need not be integers, although that is a very common case. For a (univariate) continuous random variable, all we can say is that the random values lie in some interval of real numbers, but we cannot list all the theoretical possibilities. For example, the height of a randomly selected adult person is a random variable X whose value lies in some interval, say 36 inches to 96 inches. Of course when we actually measure the height of an individual we do not obtain arbitrary values in this range, since our measuring instruments will limit the precision to say the nearest 0.1-inch. However, as a theoretical description, and remember that's what a random variable is supposed to give, it turns out to be mathematically simpler to assume that the measurements have infinite precision and therefore can take any value in the specified interval. We now turn our attention to making these ideas more precise. This chapter will focus on discrete random variables.

13.2 Discrete Random Variables

We first give a definition that makes precise the discussion in section 13.1. This will be followed by a number of examples, which should give the reader a more concrete grasp of the concept. In section 13.3 we will consider some special discrete random variables that have wide applicability.

Definition 13.1: A *discrete random variable* associated with an experiment is a set of numerical outcomes $x_1, x_2, \ldots, x_n, \ldots$ such that whenever the experiment is performed one and only one of the outcomes $x_1, x_2, \ldots, x_n, \ldots$ occurs.■

We will denote random variables using capital letters, for example X or Y. A particular value of the random variable will be denoted using a lower case letter, say x or x_1, etc.

Example 13.1: Examples of discrete random variables.

a) If a single die is tossed, the number of dots on the upward facing side defines a random variable X whose values are $\{1,2,3,4,5,6\}$. If we toss a pair of dice and record the sum Y, that is again a random variable whose values are $\{2,3,4,5,6,7,8,9,10,11,12\}$.

b) If we take a survey of 25 people to find whether they support candidate A, then the number of supporters of A in the sample surveyed is a random variable X whose values are $\{0,1,2,\ldots,24,25\}$. If we are interested in the fraction of people in the survey that support candidate A, rather than the number, this is again a discrete random variable Y taking values $\{0,.04,.08,\ldots,.96,1\}$, obtained by dividing the actual number X by 25.

c) If we count the number of cars going through a tollbooth during the same one-minute interval on various days we have a random variable X whose possible values are $\{0,1,2,\ldots\}$. In this example, one could reasonably argue that a value of say 1000 is impossible. However, rather than setting an arbitrary cutoff we will deal with this objection by assigning extremely small probabilities to such exceptional values.■

So far a random variable only describes what are the possible outcomes of an experiment. To be useful we must give some information regarding the likelihood of each value.

Definition 13.2: The *probability distribution* of a discrete random variable is the set of probabilities $p_1, p_2, \ldots p_n, \ldots$ associated with the outcomes $x_1, x_2, \ldots, x_n, \ldots$.■

For a discrete random variable, it is often convenient to exhibit the values and their associated probabilities in a table, which is often referred to as the probability distribution.

Example 13.2: Find the probability distributions for the random variables in Example 13.1.

Solution:

The probability distributions for the random variables X and Y in Example 13.1a) are easily obtained from our work with probabilities in Chapters 10 and 11. Namely, for tossing a single die we have:

X	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

The first row of the probability distribution lists the values of the random variable. The second row lists the associated probabilities. The lead entry in the second row, $P(X = x)$, can be read as "the probability that the random variable X takes on the value x". In this example the probabilities associated with the six values of the random variable are all 1/6, since any side has equal chance of facing upwards.

For the sum of two dice Y we have the following probability distribution.

Y	2	3	4	5	6	7
$P(Y = y)$	1/36	2/36	3/36	4/36	5/36	6/36
Y	8	9	10	11	12	
$P(Y = y)$	5/36	4/36	3/36	2/36	1/36	

Table 13.1

The entry under $Y = 4$ for instance asserts that $P(Y = 4)$ is 3/36. This result is obtained from our basic probability analysis of this experiment, as described in Example 10.6 of Chapter 10. There are three toss combinations $(1,3),(2,2),(3,1)$ out of 36 possible outcomes that produce a sum of the dice equal to 4.

What are the probability distributions for parts b) and c) of Example 13.1? This cannot be answered without further knowledge of the underlying processes. As we will see in section 13.3 the distribution for b) depends on the percent of voters favoring candidate A, while in c) the probabilities depend on the average number of vehicles crossing during the particular one-minute time period.■

In each of the tables given above, the sum of the probabilities totals one. This is because our definition of random variable required that whenever the experiment is performed one and only one value of the random variable must be obtained. Thus the events $\mathrm{E}_1 = \{\ X$ has value $x_1\ \}$, $\mathrm{E}_2 = \{\ X$ has value $x_2\ \}, \ldots\ \mathrm{E}_n = \{\ X$ has value $x_n, \ldots\ \}$ are mutually exclusive and exactly one of them is certain to occur. Thus

$$1 = P(\mathrm{E}_1 \text{ or } \mathrm{E}_2 \text{ or} \ldots \text{or } \mathrm{E}_n \text{ or} \ldots) = P(\mathrm{E}_1) + P(\mathrm{E}_2) + \cdots + P(\mathrm{E}_n) + \cdots = p_1 + p_2 + \cdots + p_n + \cdots.$$

We summarize this as a rule.

Rule 13.1 (Unity Rule): If X is a discrete random variable with values $x_1, x_2, \ldots x_n, \ldots$ and associated probabilities $p_1, p_2, \ldots p_n, \ldots$, with $p_i \geq 0$ then $p_1 + p_2 + \cdots + p_n + \cdots = 1$.■

Rather than give the probability distribution in tabular form it is visually more appealing to present it graphically. The resulting graph is called the *probability histogram* of the discrete random variable.

Definition 13.3: The *probability histogram* of the random variable X whose probability distribution is given by the table

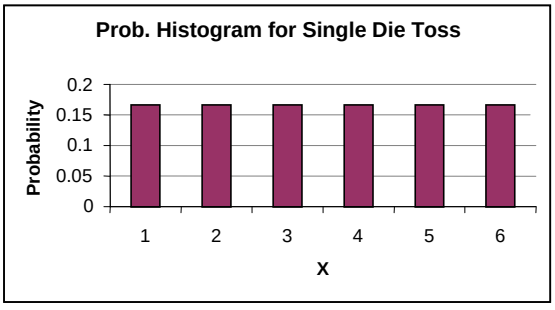
X	x_1	x_2	…	x_n	…
$P(X = x)$	p_1	p_2	…	p_n	…

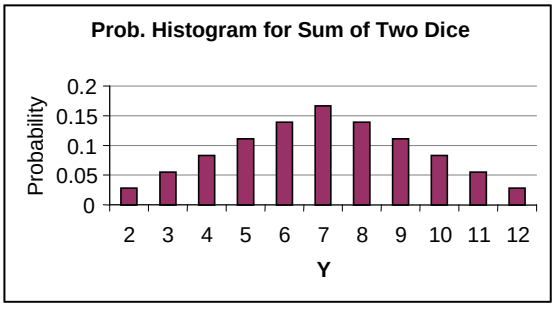
is a column chart whose horizontal axis shows the values $x_1, x_2 \ldots, x_n$ arranged in increasing order, with a thin bar centered around each value x_i and having height equal to the corresponding probability p_i.■

Example 13.3: Construct the probability histograms for the random variables X and Y in Example 13.2.

Solution:

The probability histograms for the random variables X and Y in Example 13.2 are given below. The width of the bars is somewhat arbitrary and depends on the number of bars that have to be shown and the space available for the display. If the random variable assumes many values, one often simply replaces the bar with a dot plotted over the point x_i at height p_i. The tech notes at the end of the chapter give details on using *Excel* to draw a probability histogram.





■

Random variables are supposed to give a theoretical model for the outcomes of an experiment. How well do the models for dice tossing described by the random variables X and Y account for empirical data? We can perform a simulation to make an assessment. The technical details for

how to do this in *Excel* are described in the tech notes. Chapter 14 goes a little more deeply into the theory behind these simulations.

Example 13.4: Simulate the tossing a single die 500 times. Compare the results with the theoretical probability distribution.

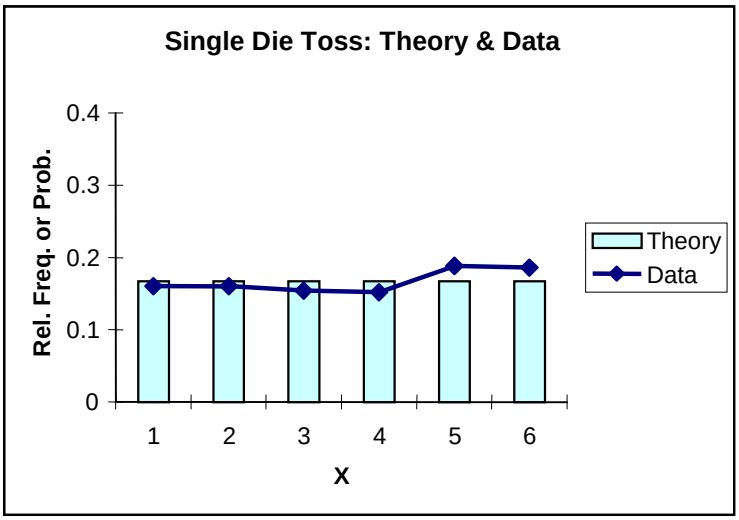
Solution:

The table below shows the results of a simulation of 500 tosses of a single die. The relative frequencies in row 3 are obtained from the actual frequencies in row 2 by dividing each entry in the latter row by 500.

Outcome, x	1	2	3	4	5	6
Simulation Freq.	80	80	77	76	94	93
Simulation Rel. Freq.	.16	.16	.154	.152	.188	.186
$P(X = x)$	.166	.166	.166	.166	.166	.166

Table 13.2

There is generally good agreement between the empirical relative frequencies in row 3 and the theoretical probabilities in row 4. A statistical procedure known as the chi-square test may be used to show that the deviations exhibited are well within the expected variations for this type of experiment. The empirical relative frequencies and the theoretical probabilities may also be compared graphically as in the figure below. The bars are used to represent the theoretical probability histogram and the connected points show the relative frequency for the data.



Similar graphs can be obtained using the simulation program *dice.xls*, except that the empirical frequencies are plotted with bars and the theoretical values, after the user enters them, are plotted as a connected line.■

As was mentioned in the introduction, we can define numerical measures associated with a random variable that capture the flavor of the mean and standard deviation for data. Consider the experiment that is summarized in Table 13.2. What is the average of the 500 tosses? To obtain this we must add the values found in all 500 tosses and then divide by 500. The sum of the tosses can be obtained using the information in rows 1 and 2. Namely,

$$\text{sum} = (1\times 80) + (2\times 80) + (3\times 77) + (4\times 76) + (5\times 94) + (6\times 93) = 1803 \qquad (13.1)$$

Note the similarity with the method used in section 7.6 on estimating the mean and standard deviation. In this case though the data is naturally grouped into the values 1, 2, 3, etc. and the above calculation is not an approximation but the exact value of the sum of the tosses. To obtain the mean of the tosses we divide the sum by 500, the number of tosses, obtaining 3.606. For our immediate purposes it is more important to examine the form of the answer, rather than its numerical value. In computing the value of *sum/ 500* we can divide each term on the right side of (13.1) by the denominator 500, obtaining

$$\text{mean} = \frac{\text{sum}}{500} = 1\times\frac{80}{500} + 2\times\frac{80}{500} + 3\times\frac{77}{500} + 4\times\frac{76}{500} + 5\times\frac{94}{500} + 6\times\frac{93}{500}$$

$$= (1\times 0.16) + (2\times 0.16) + (3\times 0.154) + (4\times 0.152) + (5\times 0.188) + (6\times 0.186).$$

In the latter form for the mean, each value 1 to 6 is weighted with the relative frequency with which it occurred in the experiment. Since these relative frequencies are approximations to the theoretical probabilities (compare rows 3 and 4 in Table 13.2) we are lead to the following definition for the theoretical mean of a discrete random variable.

Definition 13.4: The *expected value* or *mean* of a random variable X having a discrete probability distribution given by the table

X	x_1	x_2	…	x_n	…
$P(X=x)$	p_1	p_2	…	p_n	…

is the quantity denoted by $E(X)$ or μ_X and defined as

$$E(X) = \mu_X = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n + \cdots .\blacksquare$$

Remark: μ is the Greek letter ***m****u*, which of course should bring to mind the analogous **m**ean of a data set. ***Do not make the mistake of dividing the sum given in*** Definition 13.4 ***by the number of values of the random variable***. For a random variable, the averaging is accomplished with the weighting factors p_i, each of which accounts for the frequency with which a value x_i is supposed to occur.

Example 13.5: Find the expected value for the toss of a single die and for the sum of the tosses two dice.

Solution:

The expected value of the random variable X given by the toss of a single die is

$$1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=\frac{1}{6}\times(1+2+3+4+5+6)=\frac{21}{6}=3.5.$$

Note that this answer is close to the average 3.6 observed in the simulation of Example 13.4. As the number of tosses increases the average for an actual sample is very likely to get close to the theoretical average given by $E(X)$. This important point will be explored further in Chapter 15.

For the random variable Y that gives the sum of two tosses, a calculation using the probability distribution in Table 13.1 yields that $E(Y)=7$. Note that this is twice the average or expected value for the toss of a single die. This is not surprising, since Y is just the sum of the values obtained on each individual die.■

Example 13.6: A box contains 50 balls, of which 10 are black and 40 are red. Consider playing the following game. For the price of \$1 you are blindfolded and allowed to reach into the box to draw a ball. If the ball is black you receive back \$4; if it is red you get back nothing and thus lose the \$1 you paid to play. If you play this game repeatedly, on average how much will you win or lose?

Solution:

The question involves a calculation of a theoretical average. We can do this by constructing a random variable that describes the possible outcomes of playing the game. Let us call this random variable X. We need to list the values of the random variable and then the probabilities associated with each. There are two values for X. If you draw a black ball you will have made \$3 (the difference between what you get back and the fee to play). If you draw a red ball you will simply lose \$1. We can represent this as a value of -1 for X. Since the probability of selecting a red ball is 40/50 = 4/5 and the probability of selecting a black ball is 10/50 = 1/5 we have the following probability distribution for X.

X	-1	3
$P(X=x)$	4/5	1/5

Thus $\mu_X=(-1)\times\frac{4}{5}+3\times\frac{1}{5}=-\frac{1}{5}$, so that, on average, you will lose 1/5 of a dollar, or 20¢ each play. In the tech notes we will see how this experiment may be simulated and the theoretical average approximated by the data average.■

In summarizing data we needed, in addition to the average, a measure of how the data values are spread around the average. We want a similar measure for random variables.

Definition 13.5: The *standard deviation* of a random variable X from its mean μ is the number denoted by σ_X or just σ and defined by

$$\sigma = \sqrt{(x_1-\mu)^2 p_1 + (x_2-\mu)^2 p_2 + \cdots + (x_n-\mu)^2 p_n + \cdots}.$$

The *variance,* var_X, of a random variable X is the square of the standard deviation, i.e. $\text{var}_X = \sigma_X^2$.■

The symbol σ is the Greek letter ***s****igma*, meant to suggest the Latin letter ***s*** used to represent the standard deviation of a set of data. The definition of σ has the same spirit as the definition of μ. Both are weighted averages. In σ, we square the deviation of each value x_i from the mean value μ. We weight each squared deviation with the frequency or probability p_i with which the corresponding value x_i occurs. The outer square root has the same purpose as it does for the standard deviation of data, namely, to express the answer in similar units as the quantities x_i.

Example 13.7: Compute the standard deviation for the random variables X and Y in Example 13.2.

Solution:

We have for X,

$$\sigma_X = \sqrt{(1-3.5)^2\frac{1}{6} + (2-3.5)^2\frac{1}{6} + (3-3.5)^2\frac{1}{6} + (4-3.5)^2\frac{1}{6} + (5-3.5)^2\frac{1}{6} + (6-3.5)^2\frac{1}{6}} \approx 1.71$$

The terms under the radical give the variance. Its exact value is $\text{var}_X = 17.5/6 \approx 2.92$. For the simulated dice tossing experiment given in Table 13.2, the standard deviation of the data set is 1.74, which compares favorably with the theoretical value of σ_X computed above.

For the sum of two dice, using a similar computation with Table 13.1, we find that $\sigma_Y \approx 2.42$ and $\text{var}_Y = \sigma_Y^2 \approx (2.42)^2 \approx 5.86$. Note that though the standard deviation of the random variable Y is greater than that of X, it is not twice as great, as was the relationship between the expected values. Remarkably though, the variance of Y is precisely double the variance of X, although because of round off this is not exactly realized in the above computations. From the relationship $\sigma_Y^2 = 2\sigma_X^2$ we obtain that $\sigma_Y = \sqrt{2}\,\sigma_X$. We will return to this issue in Chapter 15.■

We will get a better idea of the significance of the standard deviation in the next section as well as in our discussion of the normal distribution in Chapter 14.

13.3 The Binomial Distribution

Example 13.8: Four unrelated people walk into a blood donation center. Suppose X is the number of these who are of type A. Find the probability distribution of X.

Solution:

Clearly the possible values of X are 0, 1, 2, 3, 4. The difficulty is in computing the probability of each of these outcomes. The extreme possibilities are easy. $X = 0$ means that none of the donors was of type A. Based on the data in Chapter 11, Example 11.1, we know that $P(A) = .4$ and therefore that $P(A^c) = .6$. Since the donors are unrelated, the events that none of them are of type A are independent and so the probability that $X = 0$ is given by $(.6)^4 \approx .13$. Similarly, $P(X = 4) = (.4)^4 \approx .026$.

To find $P(X = 1)$ we must examine the possible ways this event can occur. To simplify the notation let us use S (for success) when a donor has type A, and F (for failure) otherwise. The random variable X will have value 1 when exactly one out of the four donors has type A. There are four possible ways in which this can happen: SFFF or FSFF or FFSF or FFFS. Using independence, the probability of the sequence SFFF is $(.40)(.60)^3 \approx .086$. The other three arrangements also have the same probability and since they are mutually exclusive we can add the probabilities to get the probability of SFFF or FSFF or FFSF or FFFS. Thus $P(X = 1) = 4(.40)(.60)^3 \approx .346$.

To obtain $P(X = 2)$ we list all possible ways in which exactly two out of the four donors could be of type A. We obtain the following list: SSFF or SFSF or SFFS or FSSF or FSFS or FFSS. Any one of these six outcomes has the same chance of occurring, which independence gives as $(.40)^2(.60)^2$. Thus $P(X = 2) = 6(.40)^2(.60)^2 \approx .346$.

Finally, the reader should check that $P(X = 3) = 4(.40)^3(.60) \approx .154$. The results are summarized in tabular form below.

X	0	1	2	3	4
$P(X = x)$	.130	.346	.346	.154	.026

The sum of the probabilities in the second row is 1.002 instead of the theoretically correct value of 1 as stated in the Unity Rule. This is caused by round-off errors in the decimal values listed in the table.■

Example 13.8 illustrates a random variable of the binomial type. Such a variable arises in the following situation. We perform an experiment in which we focus on one particular outcome that we consider a success, and which we label S. Any other outcome of the experiment is considered a failure, denoted by F. The experiment is performed n times in such a way that the outcomes on

each trial are independent of each other. The quantity X giving the number of successes obtained in the n trials is called a *binomial random variable*. The following rule gives the formula for the probability distribution of such a random variable.

Rule 13.2 (Binomial Distribution): Suppose a certain outcome S (success) of an experiment has a probability p of occurring whenever the experiment is carried out. The complement of S, denoted by F, will then occur with probability $q = 1 - p$. Let X denote the number of successes obtained when the experiment is repeated independently n times. X is a discrete random variable. Its probability distribution is given by

X	0	1	2	…	k	…	n
$P(X=k)$	q^n	${}_nC_1 pq^{n-1}$	${}_nC_2 p^2 q^{n-2}$	…	${}_nC_k p^k q^{n-k}$	…	p^n

The numbers ${}_nC_k$ appearing as coefficients in row 2 are called binomial coefficients and are given by the formula

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

- We have deviated slightly from our usual convention of denoting generic values of a random variable X with the letter x. Here we denote an unspecified value of X by the letter k to emphasize that we are dealing with integer values.

- Since we are performing the experiment n times the number of successes must certainly be one of the numbers 0, 1, 2, … n, as listed in the table.

- The event that $X = 0$ means we had no successes in the n trials so that all n outcomes were failures. Thus, using independence of the trials we have $P(X = 0) = P(\text{F and F … and F}) = q^n$ as stated in the second row. The probability for all successes, $P(X = n) = p^n$, can be obtained by similar reasoning. These two special cases should be available at your fingertips, so it is good to understand them independently of the general result.

- The symbol $k!$ is read "k factorial." We remind the reader that for $k \geq 1$ we have $k! = k \times (k-1) \times \cdots \times 2 \times 1$. In order that the general formula remain valid for $k = 0$ and $k = n$ we conventionally define $0! = 1$. The binomial coefficients ${}_nC_k$ count the number of ways in which k successes can appear in the sequence of n trials. For example,

$$ {}_4C_2 = \frac{4!}{2!2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1) \times (2 \times 1)} = \frac{4 \times 3}{2 \times 1} = 6, $$

  which the reader can check in Example 13.8 was the number of ways that two suitable donors could appear in the group of four. In computing binomial coefficients by hand one should first

perform the numerous cancellations of common factors in the numerator and denominator. Thus

$$ {}_{40}C_6 = \frac{40!}{34!\,6!} = \frac{40(39)(38)(37)(36)(35)\cancel{(34!)}}{\cancel{34!}\;6!} = \frac{(\overset{5}{\cancel{40}})(\overset{13}{\cancel{39}})(38)(37)(\overset{6}{\cancel{36}})(\overset{7}{\cancel{35}})}{\cancel{6}(\cancel{5})(\cancel{4})(\cancel{3})(\cancel{2})} = 3,838,380 $$

- The independence of the trials implies that a particular sequence of n trials that produces k successes has a probability of occurring given by $p^k(1-p)^{n-k} = p^k q^{n-k}$. The previous paragraph tells us that there are ${}_nC_k$ such outcomes and this implies the general probability formula stated in the table. Notice the structure of the formula. The probability p of success is raised to the power k, which is the number of successes we are interested in. The exponent of q, the failure probability, is then the number of failures $n-k$ in the sequence of n trials.

Example 13.9: Compute the probabilities in Example 13.8 using the formula for the binomial distribution.

Solution:

Using the binomial distribution requires one to identify two "parameters", namely the probability p of success and the number n of trials. (It is also important to make sure that the trials are independent.) Here $p = .4$ and therefore $q = 1-p = .6$ and $n = 4$. To three decimal places, the probabilities are given by:

$$P(X=0) = .6^4 = .130$$

$$P(X=1) = {}_4C_1(.4)^1(.6)^3 = \frac{4!}{1!3!}(.4)(.6)^3 = \frac{4\times3\times2\times1}{3\times2\times1}(.4)(.6)^3 = 4(.4)(.6)^3 = .346$$

$$P(X=2) = {}_4C_2(.4)^2(.6)^2 = \frac{4!}{2!2!}(.4)^2(.6)^2 = \frac{4\times3\times2\times1}{(2\times1)(2\times1)}(.4)^2(.6)^2 = 6(.4)^2(.6)^2 = .346$$

$$P(X=3) = {}_4C_3(.4)^3(.6)^1 = \frac{4!}{3!1!}(.4)^3(.6)^1 = \frac{4\times3\times2\times1}{3\times2\times1}(.4)^3(.6) = 4(.4)^3(.6) = .154$$

$$P(X=4) = (.4)^4 = .026 \;\blacksquare$$

The computation of binomial coefficients is quite formidable, although many calculators now have this capability built-in. In the tech notes section we indicate how this may be done in *Excel*. Appendix B gives three place tables of the binomial distribution for various probabilities of success and $n = 2,3,\ldots 15$. These can be used as a convenient substitute for the formula or a statistical program.

Example 13.10: A coin is tossed 10 times. Using the tables find the probability that the number of heads obtained will be 4, 5 or 6.

Solution:

The value of $p = .5$ and we are performing $n = 10$ trials. In appendix section B.1 find the entry for $n = 10$ and then look at the column headed with .50 at the top. The probability of getting exactly 4 heads is listed in the column headed with a k. The entry for $k = 4$ (successes) is .205, for 5 successes is .246 and for 6 successes is .205. The probability of getting 4 or 5 or 6 heads is the sum of the three probabilities, which is .656.■

When the number of trials is large we are often more interested in knowing the probability that the number of successes falls in a certain range, rather than the individual probabilities. In this case it is more useful to have a table of cumulative probabilities. Symbolically, this is the probability

$$P(X \le k) = P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = k).$$

The tables in section B.2 give these cumulative probabilities for $n = 20, 25$ and 30. Let's see how they are used.

Example 13.11: A certain standardized exam has a pass rate of .70, i.e. 70 % of students who take the exam pass it. If the exam is administered to 25 randomly selected students determine the following probabilities.

a) The probability that 15 or fewer students will pass.

b) The probability that 20 or more will pass.

c) The probability that the number of students who pass, X, will satisfy $15 \le X \le 20$.

a) We need $P(X \le 15)$. The table entry for $k = 15$ gives this value as .189. Therefore, it is very likely that the number of passing students will exceed 15.

b) The table gives the probabilities up to a certain point, whereas this question asks for the probability beyond a certain point. However, we can use complementation to express the latter probability in terms of cumulative ones. Namely,

$$P(X \ge 20) = 1 - P(X < 20) = 1 - P(X \le 19) = 1 - .807 = .193.$$

Notice that the complement of getting 20 or more passes is getting <u>strictly fewer</u> than 20. Since the number of successes can only be an integer this is equivalent to $X \le 19$.

c) We want $P(15 \le X \le 20) = P(X = 15) + P(X = 16) + \cdots + P(X = 20)$. We can obtain this from the cumulative probability through 20 by removing the contribution of all terms where X is smaller than 15. In other words, we have

$$P(15 \le X \le 20) = P(X \le 20) - P(X < 15) = P(X \le 20) - P(X \le 14) = .910 - .098 = .812 \blacksquare$$

One might ask why we don't include any tables for the binomial distribution beyond $n = 30$. Of course one obvious answer is that such tables would take up too much space. This is quite true,

but a more important reason is that such tables are not necessary. As we will see in the next chapter, when n is larger than 30 (and as long as p is neither too close to zero or one) the binomial distribution can be approximated quite easily using the normal or bell-shaped distribution that we have referred to earlier in the course.

At this point it is instructive to take a look at the histogram of the binomial distribution. The file *distributions.xls* allows one to easily examine these histograms as well as the numerical values of the binomial distribution probabilities and cumulative probabilities.

Example 13.12: Construct histograms for binomial distributions with $p = 0.30$ and $n = 10,\ 20,\ 40,\ 80$.

Solution:

The histograms are shown below. Probability values that are too small are not visible in the pictures. There is a general drift to the right of the most probable values, although notice that the scale on the probability or vertical axis is shrinking as n increases. When $n = 80$ no specific number of successes in the 80 trials has a probability of occurrence exceeding 0.1, while for $n = 10$, there is a greater than 25% chance of getting exactly 3 successes. Overall, the histograms become more symmetric in appearance and bell-shaped, although a close examination of the graphs will reveal that the symmetry is only approximate. It is only when $p = .5$ that the histogram has exact symmetry around the peak point.

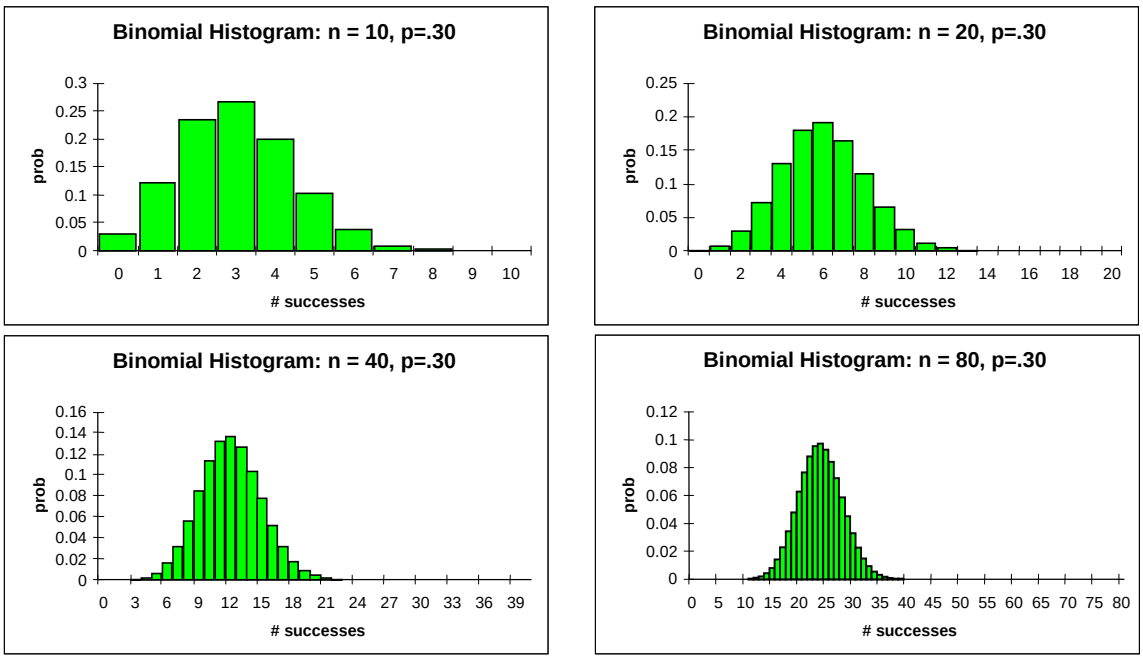


Figure 13.1

■

In our later work we will need to know the mean and standard deviation of the binomial distribution. These can be derived from the general formulas for the distribution, but we will spare the reader the details.

Rule 13.3: If X has a binomial distribution with n trials and probability of success p then $\mu_X = np$, $\sigma_X = \sqrt{npq}$ and the variance is $\text{var}_X = npq$.■

The formula for the mean μ is quite plausible. For example, if $n = 40$ and $p = .3$ as in Example 13.12, we would have $\mu = 40 \times .3 = 12$, so that on average we would have 12 successes. Notice that the histogram in Figure 13.1 for $n = 40$ reaches its peak at approximately $n = 12$ and similarly for the other histograms the peak occurs at approximately np. Thus np is also the most probable number of successes, though as remarked above the probability of hitting this value exactly becomes quite small when the number of trials n is large.

The formula for the standard deviation is not intuitively obvious. However, it quantifies an aspect of the histograms that is very important to appreciate and which we will use extensively in statistical inference. In the histograms in Figure 13.1 the values plotted along the horizontal axis double as we double n. This is not surprising, since this axis represents the number of possible successes and we are doubling that from $n = 10$ to 20 up to $n = 80$. Of course as n increases most of the probabilities for the number of successes are too small to show on the graph, which becomes concentrated on a steadily smaller fraction of the horizontal axis. As we will show later, this phenomenon is driven by the formula for the standard deviation and in particular that as n increases, σ only grows like the $\sqrt{n}$ rather than like n.

Example 13.13: Compute the mean and standard deviation for the binomial distributions exhibited in Figure 13.1.

Solution:

In each case we have $p = .3$ and therefore $q = .7$. The formulas in Rule 13.3 then give the following values for the mean and standard deviation.

n	10	20	40	80
μ	$10 \times .3 = 3$	$20 \times .3 = 6$	$40 \times .3 = 12$	$80 \times .3 = 24$
σ	$\sqrt{10 \times .3 \times .7} \approx 1.45$	$\sqrt{20 \times .3 \times .7} \approx 2.05$	$\sqrt{40 \times .3 \times .7} \approx 2.90$	$\sqrt{80 \times .3 \times .7} \approx 4.10$

The means double as we double n, but the standard deviation only increases by a factor of $\sqrt{2}$ with each doubling of n. Thus although the spread (as measured by σ) is increasing it does so much more slowly than does the entire range of values for X. As a result the histogram becomes more concentrated in appearance as n increases.■

13.4 The Poisson Distribution

The *Poisson* (pronounced **Pwah**-sone) *Distribution* is named after the mathematician Siméon-Denis Poisson who discovered it in the 1830s. Mathematically this distribution approximates the binomial distribution when the value of p is small and the value of n is large. For example, $p = 0.02$ and $n = 150$. This approximation was useful in the pre-computer world, since the probabilities for the Poisson distribution are easier to calculate than the formulas for the binomial distribution. As a practical matter this particular use of the Poisson distribution is now of little importance. However, for reasons we will see in the next chapter, the distribution arises in other contexts that make it a very useful model for certain types of data. For example, the situation described in Example 13.1c) involving cars passing through a tollbooth can often be modeled using a Poisson distribution.

Unlike the other discrete random variables we have considered in this chapter, the Poisson distribution can take on infinitely many values, 0, 1, 2, … . The probability distribution depends on one parameter, usually denoted by the Greek letter λ, (lambda). The meaning of this parameter will be discussed later, but for now note that λ can be any positive real number and is not restricted to being an integer.

Definition 13.6 (Poisson Distribution): A random variable X has a Poisson distribution with parameter λ if X can take on any of the values 0, 1, 2, … with probabilities given by the following distribution table:

X	0	1	2	…	k	…
$P(X = k)$	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2 e^{-\lambda}}{2!}$	…	$\frac{\lambda^k e^{-\lambda}}{k!}$	…

■

Example 13.14: Construct a probability table for a Poisson distribution with $\lambda = 1.5$. Display the probabilities to three decimal places.

Solution:

Using $\lambda = 1.5$ in Definition 13.6 we have

X	0	1	2	3	4	5	6	7
$P(X = k)$	0.223	0.335	0.251	0.126	0.047	0.014	0.004	0.001

The 8 probabilities shown in the table sum to one, at least to the displayed accuracy. Of course this is not exactly correct as in fact the remaining probabilities are non-zero but very small. For example,

$$P(X=8)=\frac{1.5^8 e^{-1.5}}{8!}\approx .00014\,.$$ ■

The parameter λ that appears in the definition of the Poisson distribution has a very simple interpretation, as stated in the next theorem. We omit the proof of this result, which involves some properties of the exponential function that are beyond the prerequisites for the course.

Rule 13.4: If X has a Poisson distribution with parameter λ then we have $\mu_X=\lambda$, $\sigma_X=\sqrt{\lambda}$ and $\text{var}_X=\lambda$. ■

Thus the quantity λ is the average or mean value of the random variable. This fact is important in most applications of the Poisson distribution, in particular the application to the binomial distribution referred to earlier. The idea is to compare a binomial distribution having parameters p and n with a Poisson distribution with parameter $\lambda=np$. In doing this we insure that the two distributions have the same mean, which would certainly be a requirement if they were to approximate each other. Using the file *distributions.xls* we can compare the two distributions and judge the adequacy of the approximation.

Example 13.15: Compare the binomial distributions having parameters

a) $p=0.02$ and $n=150$

b) $p=0.2$ and $n=150$

with suitable Poisson distributions.

a) According to the preceding discussion we should use a Poisson distribution with $\lambda=np=150\times 0.02=3$. Figure 13.2 below shows the histogram of the binomial distribution and a superimposed line graph that plots the Poisson probabilities ($\lambda=3$) for the same values of each random variable. There is a very close agreement between the two, and the numerical table given below the histogram confirms this.

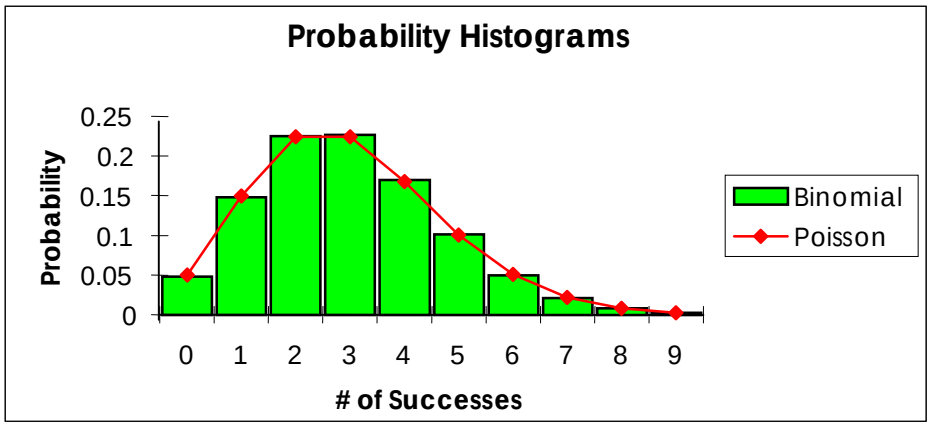


Figure 13.2

k	0	1	2	3	4	5	6	7	8	9
Binomial Prob.	0.048	0.148	0.225	0.226	0.170	0.101	0.050	0.021	0.008	0.002
Poisson Prob.	0.050	0.149	0.224	0.224	0.168	0.101	0.050	0.022	0.008	0.003

b) Using a Poisson distribution with $\lambda = np = 150 \times 0.02 = 3$ we do not obtain as good a fit to the binomial as in a). This is apparent in the histogram below (Figure 13.3). The numerical table (omitted) shows that while in a) the approximation is generally accurate to two significant digits (i.e. ignoring initial zeros after the decimal point), in this case we attain at most one digit of accuracy.

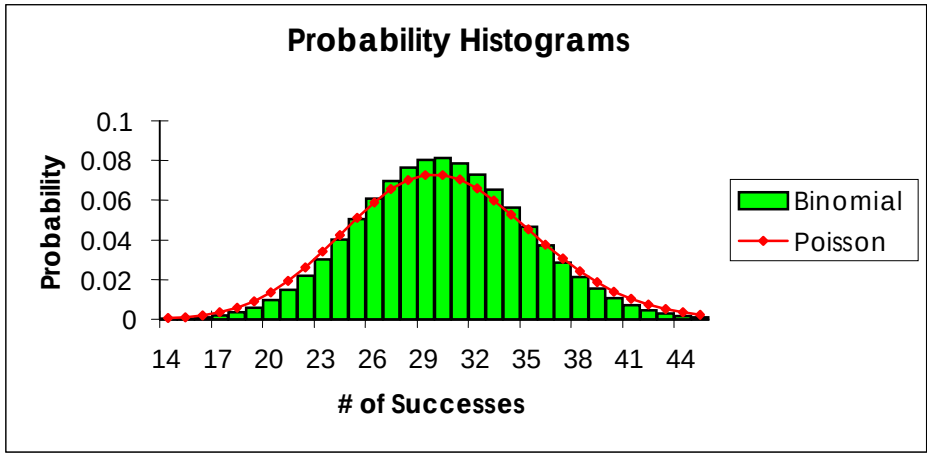


Figure 13.3

The criterion for a good fit is, roughly speaking, that n ***is large,*** p ***is small and the product*** np ***is five or less***. This is sometimes referred to as the situation when success is a rare event, but we make a large number of observations so that a few successes are likely. If n is large and the product $np > 5$ then a suitable normal distribution gives a much better approximation to the binomial distribution than does the Poisson. This will be discussed in the next chapter.■

13.5 Tech Notes

We describe implementing in *Excel* some of the procedures discussed in this chapter.

13.5.1 Plotting Probability Histograms

Example 13.16: Using *Excel* construct the probability histogram for the random variable whose probability distribution is given by

X	2	5	7
$P(X = x)$	.35	.25	.4

Solution:

To plot the histogram correctly the X values in the plot must be uniformly spaced. If this is not the case we can often arrange for it by adding extra X values to which we assign zero probability. In a blank spreadsheet enter the table as a vertical array including all integer values between 2 and 7 and assigning probability zero to the extraneous values. See the figure below. Include the text titles so you know what the numbers refer to.

X	2	3	4	5	6	7
$P(X = x)$	0.35	0	0	0.25	0	0.4

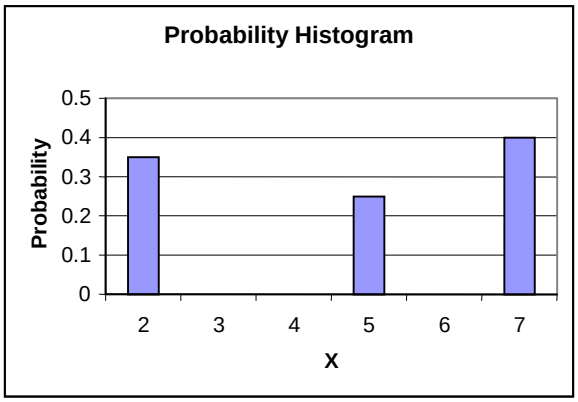
Open the chart wizard.

Step 1: Accept the default column chart type and go to step 2.

Step 2: Enter the reference or select the cells in the probability column of your table. You should see a bar chart, but the horizontal axis will not yet be correctly labeled. Clicks on the Series tab and in the box labeled "Category (X) axis labels" enter the reference or select the cells containing the X values.

Step 3: Add a title, label the horizontal axis as X or some more descriptive term. The Y axis label should be "Probability". Delete the legend; it is not needed when there is only one set of bars.

Step 4: Send the graph to the current worksheet. The final result should look as shown below. You can then modify it to suit your tastes.



■

13.5.2 Generating Random Numbers

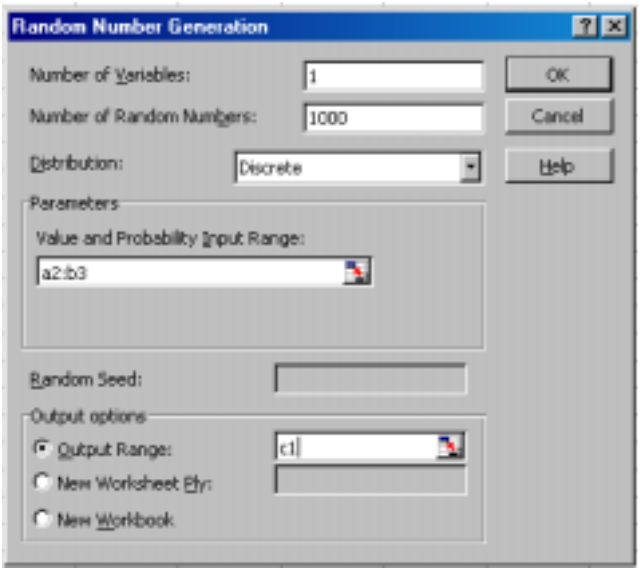
Example 13.17: Use *Excel* to simulate generating values for the random variable described in Example 13.6.

Solution:

The first step is to enter the table for the random variable as we did in the previous section. When doing a simulation it is not necessary to insert any missing values. Thus you might enter the following table in the range A1: B3

	A	B
1	X	P(X)
2	-1	0.8
3	3	0.2

Next we must generate values of this random variable, in effect simulating the play of the game. To do this click on the *Tools* menu, select *Data Analysis…* and from the choices select *Random Number Generation*. Fill in the dialog box as shown below. The entry for "Number of Variables" specifies the number of columns of random numbers you want to generate. For a simple simulation the entry here is usually set to one, but more complex simulations may require additional sets of random values. The entry for "Number of Random Numbers" gives the number of random numbers generated in each column.



The *Help* button will lead you to a clear explanation of the remaining items in the dialog box. As set up, the program generates in column C the results of "playing" the specified game 1000 times.

We can use these simulated values to confirm the calculation that on average playing this game will cost you $ 0.20. Simply use the *Stats* menu to find the numerical average of the results of these 1000 plays. This will be explored further in the exercises.■

In addition to using a discrete variable based on a table, the "Distribution" entry has a pull-down menu from which the user may select from the binomial and Poisson distributions, among others.

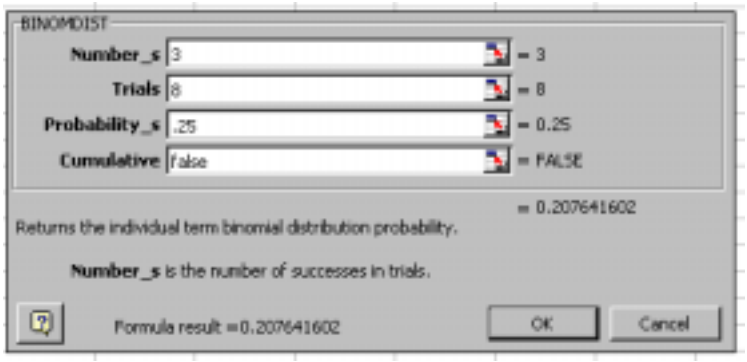
13.5.3 Computation of Binomial & Other Probabilities

Many probability distributions can be computed using built-in *Excel* functions. As it is difficult to remember the exact specifications needed to compute each function, you will want to use the function "wizard" to assist you.

Example 13.18: Use the function wizard to compute probabilities for the binomial distribution.

Solution:

Clicking on the toolbar button labeled f_x activates the wizard. Scroll down to the list of *Statistical* functions (or use the category, *Most Recently Used*, if that is relevant). To obtain probability values for a binomial distribution, select *BINOMDIST*. A dialog box opens. Most of the values that you need to enter (except for the last one) should be quite clear to you based on our discussion of the binomial distribution and the explanations in the box. For example, to compute $P(X = 3)$ if X has a binomial distribution with $p = .25$ and $n = 8$ fill out the entries as shown below:



The entry "False" for the cumulative field indicates that we wish only the specific probability of 3 successes. If we wanted the probability $P(X \le 3)$ we would enter the value "true" in this field. When you have completed entering the values click OK. The formula =BINOMDIST(3,8,.25,FALSE) will be entered in the active cell and the value .207… will appear in the spreadsheet. There is a similar function for computing values of the Poisson distribution.■

13.6 Summary

Flips of a coin, tosses of a die, arrivals and departures from a queue, may appear as simply random phenomena. However, underlying the randomness are patterns that can be described by probability theory. The notion of a ***random variable*** provides the theoretical mathematical framework for this description. A random variable has two essential features: first, a clear statement of the possible values of this random quantity, and second, a ***probability distribution*** that associates with each outcome a certain probability for its occurrence. In this chapter, we have considered discrete random variables, in which the numerical outcomes can be listed. Random

phenomena in which the possible outcomes can assume an entire interval of values will be considered in Chapter 14.

Since a random variable provides a theoretical model for random data, there are analogues for many of the constructions used in descriptive statistics. Thus, we can represent a discrete random variable graphically using a ***probability histogram***. The central tendency and spread of the distribution can be summarized through the ***mean*** (***expected value***), μ, and the ***standard deviation***, σ.

There are a number of theoretical discrete random variables that appear as models of many random experiments. We have described the ***binomial*** and ***Poisson*** distributions. The binomial distribution occurs when we count the number of times an outcome of interest to us (success) occurs in repeated independent trials of an experiment. The Poisson distribution approximates the binomial distribution when the probability of success is small, but a rather large number of trials take place. As we will elaborate in the next chapter, this random variable also appears as the natural distribution in time or space for phenomena that occur infrequently.

13.7 Exercises

1. Determine which of the following describe discrete or continuous random variables, or not random variables at all. Justify your answer. It is not necessary to give the probability distribution.
   a) The number of days in May.
   b) The number of days in a randomly selected month. (Consider only non-leap years.)
   c) The length of time it takes you to commute to school each day.
   d) The number of misspelled words on page 237 (first page of this chapter) of these notes.
   e) Your grade on the final exam in this course.

2. a) Find the missing value in the following probability distribution:

X	1	2	4	6
$P(X = x)$	0.2	0.3	.15	?

   b) For the random variable in a) what is $P(X = 1 \text{ or } X = 4)$?
   c) A random variable X takes on the values $x = 1, 2, \text{or } 3$ with probabilities given by the formula $P(X = x) = \alpha(x+1)$. Find the value of α and write the probability distribution table for X.

3. a) What is the probability distribution for the random variable in 1b)?
   b) Draw a probability histogram for the distribution you found in 3a).

4. For 52 weeks a housewife kept careful records of the number of times she went to the supermarket each week. The table below summarizes the data she collected. For example, the bolded cells indicate that on seven weeks out of the 52 weeks she made only one trip to the supermarket.

y = number of trips per week	0	**1**	2	3	4
# of weeks with y trips	1	**7**	16	19	9

   a) Using the information provided by the table construct a probability distribution table for the random variable Y = # of trips made each week.

   b) Find μ_Y. In simple terms explain the meaning of this number.

5. Suppose that X has the following probability distribution table:

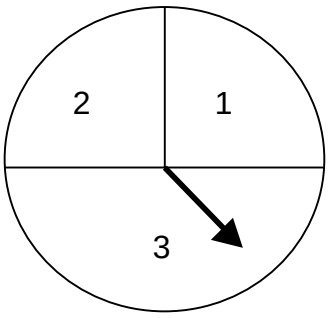
X	-2	0	2	3	5
$P(X = x)$	0.1	0.2	0.2	0.4	0.1

   a) Draw a probability histogram for X.

   b) What is $P(0 \le X \le 3)$?

   c) Find μ_X, σ_X and var_X.

6. Consider the spinner as shown below.



   Suppose you pay $1.50 to make a spin and receive back the amount in dollars of the number in the region in which the pointer lands. (Thus, as shown you would get back $3.)

   a) Write a probability distribution for your net gain when you play the game once. (Use negative numbers for a loss.)

   b) What is the average or expected gain you will have on each play?

   c) Simulate 1000 tosses of the random variable described in a). Compute the mean of the simulated data and compare with the theoretical mean value.

d) Submit to your instructor the mean of the data for your simulation in c). Construct a box plot of all the means obtained by students in the class. What is the IQR for the collection of means?

7. a) Verify the assertion made in the text that the standard deviation for the sum of two dice is approximately 2.42. (Note: You can use the symmetry of the probability distribution Table 13.1 to simplify the calculations.)

b) Use the program *dice.xls* to generate the sum of 10,000 tosses of two dice. Enter the theoretical probabilities and print out the sheet showing the frequency tally and the combined histograms.

c) From the frequency table for the 10,000 tosses that you obtained in b) use the estimation method discussed in Chapter 7 section 7.6 to approximate the mean and standard deviation of the data and compare with theoretical values for the random variable.

8. Use Binomial Distribution (Rule 13.2) to verify the values in the tables in Appendix B for the following binomial random variables.

a) $P(X = 3)$, where X has $n = 7$ and $p = .2$.

b) $P(X = 6)$, where X has $n = 10$ and $p = .7$.

9. a) If X has a binomial distribution with $n = 12$ and $p = 0.30$ find $P(2 \le X \le 6)$.

b) If X has a binomial distribution with $n = 25$ and $p = 0.30$ find $P(6 \le X \le 9)$.

10. a) If Y has a binomial distribution with $n = 25$ and $p = 0.75$, find to three decimal places $P(Y = 20)$.

b) A certain professor is known to give As to only 10% of her classes. In a class of 30 students

i) What is the expected number of As?

ii) Using a suitable table, determine the probability that there will be three or fewer As in a class of 30.

11. A bowl contains 3 black balls and 7 red ones. You select a ball from the bowl 12 times, with replacement. Let X equal the number of times a red ball is drawn.

a) Using the formula for the binomial distribution, find the probability that exactly four of the balls selected are red. Check your answer using the tables.

b) Using a table, find the probability that the number of red balls selected will be between 3 and 6, inclusive.

12. A multiple-choice test has 10 questions, each with 5 choices for the answer.

a) If you guess the answer to each question, use an appropriate table to find the probability that you will get four or more correct answers.

b) What is the expected or average number of questions you will correctly guess? Describe the statistical meaning of this answer.

13. a) A random variable X has a Poisson distribution with $\lambda = 4.5$. Compute each of the following probabilities.

i) $P(X = 2)$ ii) $P(X > 2)$ iii) $P(2 \le X \le 5)$

b) A random variable X has a binomial distribution with $n = 150$ and $p = .03$. The following table gives some probabilities for this random variable.

X	2	3	4	5	6
$P(X = k)$	0.1108	0.1691	0.1922	0.1736	0.1297
$P(Y = k)$					

Use a suitable random variable Y with a Poisson distribution to estimate the probabilities in row 2 and enter these estimates in row 3 of the table. Verify that the results provide estimates for the answers in row 2 that are correct to two significant figures.

14. Suppose that a rare illness strikes about 35 out of 10,000 children per year.

a) What is the probability that a randomly selected child will contract this illness during a single year?

b) Letting X denote the number of cases of illness among 1000 children, use a binomial distribution to fill in the missing entries in the table below.

X	0	1	2	3	4	5
$P(X = k)$						

c) What is the expected number of cases of the illness amongst a group of 1000 children?

d) Use a random variable Y with a suitable Poisson distribution to estimate the missing probabilities in b).

e) In a community with 1000 children, how likely would it be that in each of two consecutive years more than three cases of the illness would be reported? What assumptions are you using in your calculations?

15. Consider binomial distributions each having $p = .4$ with $n = 10, 20, 40, 80,$ and 160.

a) Find μ_X and σ_X for each of these.

b) Using the file *distributions.xls* or following the method discussed in section 13.5.3, fill in the probabilities in the table below: The events specified are that the value of X lies within one and two standard deviations of its mean value μ_X.

n	10	20	40	80	160
$P(\mu_X - \sigma_X \le X \le \mu_X + \sigma_X)$					
$P(\mu_X - 2\sigma_X \le X \le \mu_X + 2\sigma_X)$					

Compare the results to the Bell Curve rule for data, enunciated in Chapter 7.

16. a) A candidate believes she is favored by 60% of the voters. If this is the case, use an appropriate table to find how likely is it that a random sample of 25 voters would show fewer than 50% favoring the candidate?

b) Using *Excel's* random number generator (see section 13.5.2), generate the outcome of 1000 simulations of the experiment described in a), i.e 1000 values of a binomial random variable with $p = .60$ and $n = 25$. Determine the fraction of these experiments in which the number of successes was less than 50%. Compare with the theoretical probability found in a).

17. a) Suppose as in exercise 16 that 60% of the voters support a certain candididate. Using *Excel* construct a table giving the probability distribution for the fraction Y of voters in a survey of 25 who are in favor of the candidate. (Hint: The values of Y are 0, 1/25 = .04, 2/25 =.08, etc. The probabilities can be found by using the *binomdist* function to find the chance that there are 0, 1, 2, etc. "successes" among the 25 surveyed.)

b) Use the distribution table obtained in a) to find the mean and standard deviation of the random variable Y. Compare to the mean and standard deviation of the binomial random variable X that gives the number of voters in the surveyed sample who favor the candidate. Is there a simple relationship between the quantities μ_Y and μ_X, and the quantities σ_Y and σ_X?

c) Construct a probability histogram for the random variable Y.

18. On past final examinations a professor has observed that 70% of the students pass. The professor currently has a class of 30 students. Assume that students in the current class perform similarly to students in the past.

a) What is the probability that 25 or more students will pass the final? (Use an appropriate table.)

b) What is the probability that 15 or fewer students will pass the final?

c) What is the probability that the number of passing students will be greater than or equal to 19 and less than or equal to 24?

d) What is the expected number of students who should pass?

19. a) A computer manufacturer claims that at most 1% of the products shipped are defective. A store receives shipments of 20 computers. Assuming the worst case scenario, what fraction of shipments to the store will have at least one defective computer?

b) Using *Excel's* random number generation, simulate making 1000 shipments of 20 computers from the manufacturer described in a), assuming that there are 1% defectives in each shipment. What fraction of the 1000 shipments contain at least one defective computer? How does this number compare with the probability you computed in a)?

20. The *median* M of a random variable X is defined as a value of X such that both probabilities $P(X \le M)$ and $P(X \ge M)$ are greater than 0.5. Using the cumulative binomial distribution tables, find the median for the following binomial distributions:

a) $n = 20,\ p = .6,\ p = .2$

b) $n = 25,\ p = .6,\ p = .2$

c) Do your answers in a) and b) show any relationship to the expected value of the corresponding random variables? Based on the analogy of data and probability distributions, why might you expect this relationship to be valid?