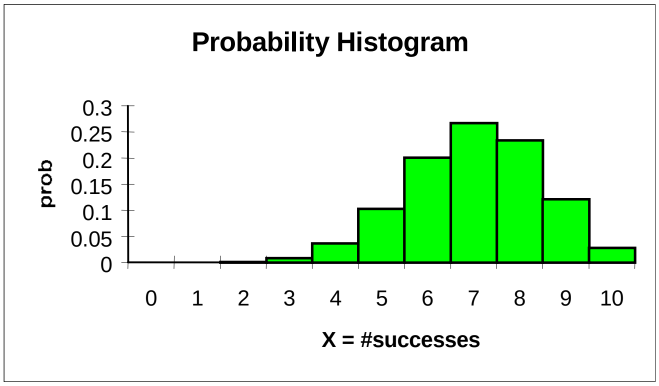
14 Random Variables II

Homer: My son, a genius!? How does it happen?
Dr.J: Well, genius, like intelligence, is usually the result of heredity and environment.
Homer: [stares blankly]
Dr.J: Although in some cases, it's a total mystery.
From: *The Simpsons*

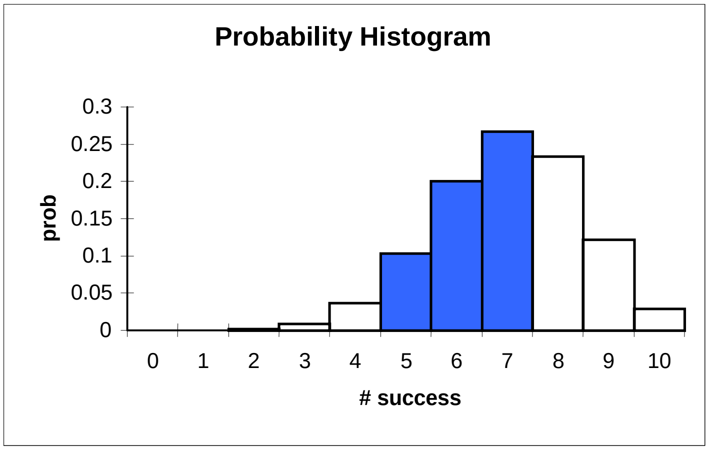
14.1 Continuous Random Variables

In the last chapter we considered discrete random variables. In order to apply probability models to a wider range of phenomena we must extend the concepts to deal with continuous random variables. The first difficulty is a fundamental one - How do we compute probabilities? For example, if X represents the length of time, in years, that a person survives after receiving a treatment for a particular type of cancer, then we can think of X as having values in the interval (0, ∞). What is the meaning of $P(X = 5.75)$? If we interpret this literally, the only reasonable answer is “zero”. Indeed, if we require that the person live exactly 5.75 years after treatment and not a moment more or less, then this is so restrictive as to be impossible. In practice, it is more reasonable to ask for the probability that a person lives between say 5.5 and 6 years. In general, if X is a continuous random variable the probability $P(X = x)$ is zero for any specific x. Non-zero probabilities may arise when we consider the probability that X falls into an interval $[a,b]$, which we write as $P(a \le X \le b)$. We need to understand how the latter probabilities are computed.

The probability histogram is the key to understanding how probabilities are computed in the continuous case. For discrete random variables we hardly gave a thought to the width of the bars used in our histograms. We will now be less casual about this and will adjust the bar width to convey some useful information. Consider first a binomial random variable X having $n = 10$ and $p = .7$. The probability histogram is shown below.



Unlike some earlier histograms for the binomial distribution, we have drawn this with no gaps between the bars. As each bar extends ½ unit to the right and left of the value of X, the width of each bar is one and therefore the area of each box is the same as the height of the box, which is the probability. Thus, in this picture we have two ways of thinking about the probability. The height or the area of each bar represents the probability of the corresponding value of X. The area representation is useful for visualizing probabilities of the form $P(a \le X \le b)$, for example $P(5 \le X \le 7) = P(X = 5) + P(X = 6) + P(X = 7)$. The latter sum is easy to visualize as the area of the three shaded rectangles shown below.



The idea of using areas to compute probabilities is the key to computing with continuous random variables. The stair-like histogram for the discrete random variable is replaced by a smooth curve and the areas under this curve give the probabilities associated with the random variable. The smooth curve defining the outline of the histogram is called the *probability density function*. Its definition is given by

Definition 14.1 (Probability Density Function) A *probability density function* of a continuous random variable X is a function $f(x)$ with the following two properties:

a) For all x, $f(x) \ge 0$.

b) The area under the graph of $f(x)$ from $-\infty$ to $+\infty$ is one.■

If these conditions hold then for any real numbers a and b, the probability that X will lie in the interval $[a,b]$, denoted by $P(a \le X \le b)$, is equal to the area under the graph of $f(x)$ over the interval $[a,b]$. The assignment of probabilities described in Definition 14.1 is often referred to as a *probability distribution* for the random variable X.

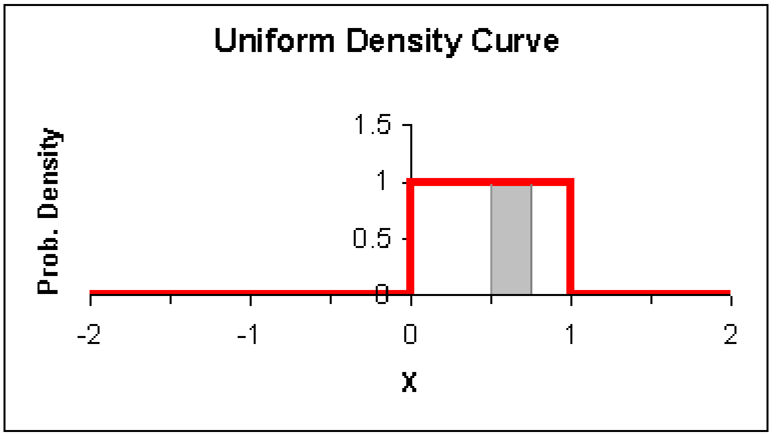
One technical point to observe is that in computing probabilities for continuous random variables it makes no difference whether we consider closed or open intervals. In other words, $P(a < X < b) = P(a \le X \le b)$. This is usually false for discrete random variables, since taking into account the possibility that $X = a$ may make a difference in the two probabilities. Indeed the event $a \le X \le b = (a < X < b)$ or $(X = a)$ or $(X = b)$. The last three events are mutually exclusive and therefore $P(a \le X \le b) = P\,(a < X < b) + P(X = a) + P(X = b)$. As we stated above, for continuous random variables $P(X = x) = 0$ and so in the continuous case we have the equality $P(a \le X \le b) = P(a < X < b)$. Similar arguments apply if only one of the $\le$ signs is replaced by strict inequality. We will often make use of this freedom in adjusting the boundaries of the interval, particularly when considering complements.

14.2 The Uniform Distribution

Our focus in this chapter will be on two continuous random variables, the uniform distribution and the normal distribution. Consider the density function given by

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{if } 1 < x \end{cases}.$$

The graph of this density function is shown below. A random variable having this density function is said to have a *uniform distribution*.



Areas under the curve represent probabilities. Thus for instance, there is zero probability of obtaining a value of X in the interval from 1 to 2 since the area under that portion of the curve is zero. On the other hand, the area of the shaded region extending from $X = 0.5$ to $X = 0.75$ is 0.25 and so $P(.5 \le X \le .75) = .25$. In general, for this random variable, if a and b are any numbers in the interval $[0,1]$ with $a < b$, then $P(a \le X \le b) = b - a$. Note that the ordinate on the above graph does <u>not</u> represent a probability, but rather a quantity called a probability density, which is the area of the region, i.e. the probability, divided by the length of the interval.

There is a simple physical model for a random variable with this density function. Suppose we aim darts at the number line, but vertical barriers prevent the darts from hitting anywhere except in the interval [0,1]. If we do not aim at any particular location in [0,1], then the chance of the dart landing in any subinterval $[a,b]$ is dependent on the length $b - a$ of that interval. There is less of a chance of hitting a small interval than a larger one. The uniform density defined above says that the chance is exactly equal to the length of the subinterval.

The uniform density plays an important role in creating computer simulations for arbitrary random variables. For example, suppose we can produce random numbers X in the interval [0,1] that are distributed according to the uniform density. If X denotes any such number, then by our discussion above $P(0 \le X \le 0.5) = 0.5$ and $P(0.5 < X \le 1) = 0.5$. Thus these two mutually exclusive events can be thought of as representing the outcomes of "Heads" and "Tails" for the toss of a coin and we can use this to construct a coin-tossing simulation. Similar, although more complicated constructions can be used to simulate other discrete and continuous random variables. See the tech notes for a more detailed description related to *Excel*. We will not describe the exact method that *Excel* and other programs use to generate these random numbers.

Many biological problems dealing with the spatial distribution of a species lead to the consideration of uniform densities in two and three spatial dimensions, and sometimes a time dimension as well (i.e. multivariate uniform distributions). ***In the biological literature uniform dispersal patterns are often referred to as random dispersion***. The picture below shows a simulation of such a pattern in a 10 by 10-rectangular region. The points were obtained mathematically by selecting the x coordinates using a uniform distribution on the interval 0 to 10 and making a similar random selection for the y coordinate. Note the occasional clustering of points. Recognizing excessive clustering patterns is often very important evidence in analyzing outbreaks of diseases or environmentally caused illness. The Poisson distribution is a prime tool in deciding whether cluster patterns deviate significantly from those produced by uniform dispersal.

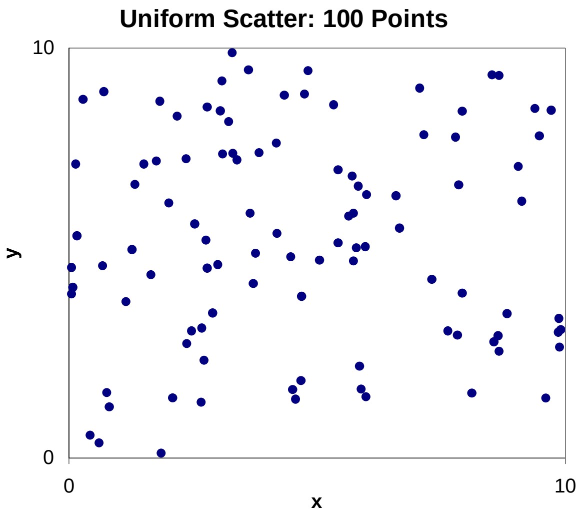


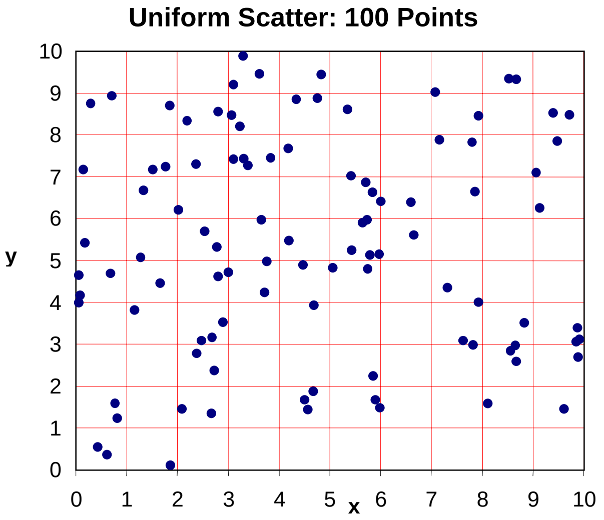
Figure 14.1

Example 14.1 (Poisson Distribution and Uniform Scatter): Suppose Figure 14.1 represents the incidence pattern of a disease in a certain geographic area. Is the observed clustering due to some important environmental factor or is this a normal artifact to be expected in random dispersal patterns?

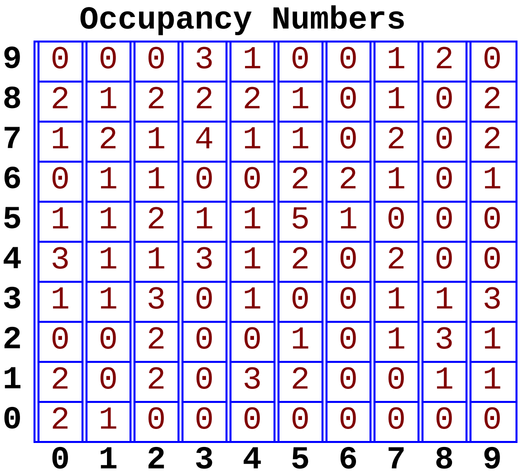
Solution:

Of course in this case we generated the data artificially from randomly scattered points, so we know the answer to the question. However, we should like to be able to carry out some analysis of the data that would give evidence of the probabilistic mechanism that produced it.

We can relate the scatter plot in Figure 14.1 to the Poisson distribution by imposing a grid over the region and counting the occupancy numbers in each square of the grid. (These square sectors are called *quadrats* in field study experiments). A superimposed grid pattern of 1×1 squares is shown below.



We then count the number of points in each square. Visually this may lead to some ambiguous cases with points that straddle the boundaries. From the mathematical perspective used in this simulation, the coordinates of each point are very exact random numbers, which almost never assume an exact integer value required for the point to land on a boundary line. In this example the occupancy number for each of the 100 squares is given in the table below.



Occupancy Numbers

9	0	0	0	3	1	0	0	1	2	0
8	2	1	2	2	2	1	0	1	0	2
7	1	2	1	4	1	1	0	2	0	2
6	0	1	1	0	0	2	2	1	0	1
5	1	1	2	1	1	5	1	0	0	0
4	3	1	1	3	1	2	0	2	0	0
3	1	1	3	0	1	0	0	1	1	3
2	0	0	2	0	0	1	0	1	3	1
1	2	0	2	0	3	2	0	0	1	1
0	2	1	0	0	0	0	0	0	0	0
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

We then make a tally of the number of cells that are unoccupied, the number of cells that have one occupant, etc. For example there is one cell with five "hits", one cell with four "hits", seven cells with three "hits". This is tabulated below.

Occupancy #	0	1	2	3	4	5
Frequency	40	32	19	7	1	1

Table 14.1

So far we have just spent a lot of effort tabulating our results. Now we assume that the points are scattered via a uniform dispersal mechanism. Since our grid contains 100 squares the chance of a point landing in a given square is .01 (all squares, having the same area, are equally likely to be hit if the dispersal is uniform). Focusing still on a particular square, the probability of k out of 100 points landing in this square is the probability that k successes will occur in 100 trials of a binomial random variable with $p = .01$. In Chapter 13, section 13.4 we have seen that a Poisson random variable may be used to approximate such a binomial distribution. The appropriate Poisson distribution will have $\lambda = np = 100(.01) = 1$. Notice that $\lambda = 1$ is also the average number of points in each 1×1 square, in the sense that we are tossing 100 points onto a grid with 100 boxes, so that each box will contain on average one point.

Using the formula for the probabilities of a Poisson distribution, we obtain the following probabilities for a square to have no "hits", one "hit" etc. These probabilities are then multiplied by 100 to obtain the expected number of squares in the grid that would have the given number of "hits". This is the third row in the table below.

Occupancy #	0	1	2	3	4	5
Predicted Poisson Frequency ($\lambda = 1$)	0.37	0.37	0.18	0.06	0.02	0.003
Predicted Frequency	37	37	18	6	2	.3

The predicted frequency pattern is then compared to the observed pattern in Table 14.1. In this case the agreement seems quite good and one might take this as reasonable evidence that the points were scattered by a uniform dispersal mechanism. In more ambiguous situations one might need to use the chi-square test, which is a statistical procedure measuring whether there is sufficient agreement between the observed and predicted values. Additional simulations of this sort may be carried out using the file *scatter.xls*.

We have here another instance of the hypothesis testing procedure mentioned earlier in Chapter 11 and which we will discuss in more detail in Chapter 16. The logic of the method is to assume the dispersal mechanism is uniform, draw from that the conclusion that the occupancy pattern will follow a Poisson distribution and then compare the data with this prediction. If the fit between model and data is good we can't be absolutely certain the dispersal pattern is uniform, but we have evidence to support it. If the fit were poor, by the standards we choose to establish, we would likely reject the hypothesis of uniform or random dispersal and look at other mechanisms that might have led to the observed pattern.

The analysis above requires a large amount of computation. A somewhat abbreviated, but less reliable alternative is described in exercise 3.■

14.3 The Standard Normal Distribution

The normal density is the most important probability density function, with the widest range of applicability. In discussing the so-called standard normal distribution it is customary to use the letter Z for the random variable and z for its values.

Definition 14.2 (The Standard Normal Density): The function $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is called the *standard normal density* function. A random variable Z has a *standard normal distribution* if for any real numbers a and b the probability $P(a \le Z \le b)$ is equal to the area under the graph of $f(z)$ between a and b.■

Because of the shape of the graph of $f(z)$ (see Figure 14.2 below) the density is often called the bell-shaped curve, although many other mathematical expressions can produce graphs with a similar appearance. As we will see, the Bell Curve Rule (Rule 7.1) derives from properties of the standard normal density. The density $f(z)$ is also called the *gaussian density*, after C. F. Gauss, the German mathematician whose investigations in the 19th century established the fundamental importance of the normal distribution.

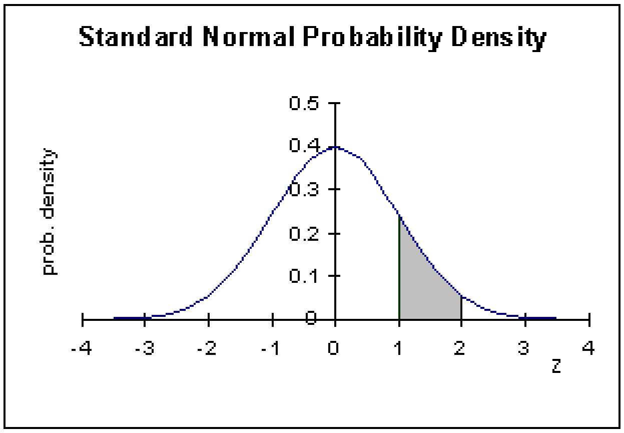


Figure 14.2

In Figure 14.2 the shaded area between one and two gives the probability that Z falls in the interval [1, 2]. Unlike the case of the uniform distribution, this area cannot be computed by simple geometric considerations. Indeed integral calculus is needed to compute the area. You will recall from calculus that areas under curves may be obtained as the value of a definite integral. In this case we have

$$P(1 \le Z \le 2) = \frac{1}{\sqrt{2\pi}} \int_1^2 e^{-z^2/2} dz .$$

Unfortunately, no matter how skillful you are at evaluating integrals the latter integral cannot be expressed in closed form in terms of the usual elementary functions. Numerical methods (Riemann sums etc.) are needed to approximate the value to any desired accuracy. From a practical viewpoint, these results have been tabulated, or a computer may be used in which a suitable routine has been provided for this computation. Here we will consider the use of tables, such as those given in section B.3. The tech notes describe the appropriate *Excel* functions.

The tables give cumulative probabilities or areas under the density curve from $-\infty$ to z. This is indicated by the schematic at the top of the table. To obtain the result for the example $P(Z < 1.65)$, we look at the first table of section B.3. In the z column locate the first two digits 1.6. The digits 0, 1, 2, … running along the top of the table are used to specify the second decimal place (except for the last row $\pm 3.$, where the top row gives the first decimal place). Scanning along the row beginning 1.6 to the column under 5, we read the entry .9505. This represents the area under the density curve to the left of $z = 1.65$ and therefore gives the probability that $Z < 1.65$. For negative values of Z we use the 2nd table of section B.3.

Before we examine some examples showing how to use the tables to compute probabilities, it is important to recall that, as for any continuous density function, the total area under the graph of $f(z)$ is one.

Example 14.2: Compute the following probabilities by finding the area of suitable regions under the normal density curve.

a) $P(Z < -.73)$

b) $P(Z > 1)$

c) $P(1 < Z < 2)$

d) $P(|Z| < 2)$

e) A value z_0 so that $P(Z > z_0) = 0.1$.

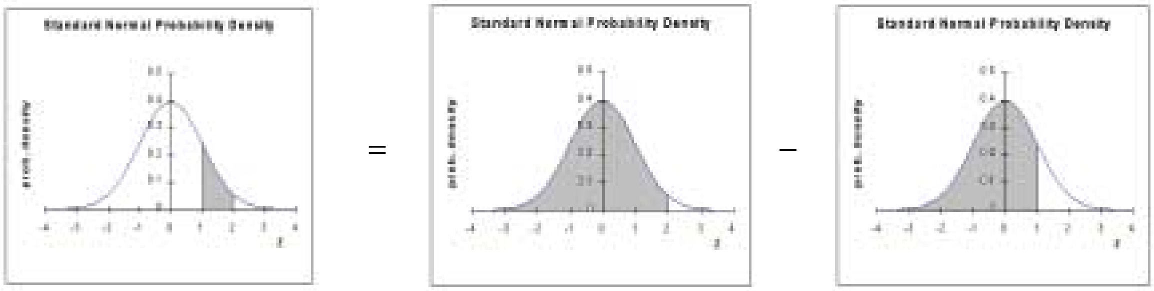
Solution:

First, note that based on our comments after Definition 14.1, we would obtain the same answers if we replaced any of the strict inequalities above with $\leq$.

a) Use the second table in section B.3 as described above. You find $P(Z < -0.73) = .2327$. Notice this is less than 1/2, as it should be, since the area involved is only a portion of the area to the left of the origin. If we had carelessly used the table for positive Z to obtain an answer of .7673 this type of simple reality check would have alerted us to an error.

b) The tables give probabilities of the form $P(Z \leq a)$. The complement of the event $P(Z > 1)$ is $P(Z \leq 1)$, which has probability 0.8413, using the entry for $z = 1.00$ in section B.3. Thus the Rule of Complements (Rule 10.5) gives $P(Z > 1) = 1 - .8413 = 0.1587$.

c) The desired probability is the area between $z = 1$ and $z = 2$. The pictogram below shows how this area can be related to the tabulated cumulative areas by appropriate subtraction.



Thus $P(1 < Z < 2) = P(Z < 2) - P(Z < 1) = .9772 - .8413 = .1359$.

d) The condition $|Z| < 2$ is the same as $-2 < Z < 2$. Using the same analysis as in c) we can evaluate the probability as

$$P(|Z| < 2) = P(-2 < Z < 2) = P(Z < 2) - P(Z < -2) = .9772 - .0228 = .9544 .$$

e) This problem involves reverse table look-up. This means we are given a probability and have to find the z value that is associated with this probability. In this case, the probability we are given is of the type $P(Z > z_0)$ and the table does not list these. However, by considering the complementary event we have that $P(Z < z_0) = 1 - 0.1 = .9$. Using the body of the table, not the z value column, we find the nearest probability listing to .9. This is found in section B.3 and corresponds to the z entry $z_0 = 1.28$, for which the listed probability is .8997.■

14.4 Normal Distributions

In the last section we considered the special standard normal distribution. In most applications more general normal distributions are needed. We describe how these are obtained and how they are related to the standard normal density.

To begin, we must say something about the mean and standard deviation for continuous random variables. We recall from chapter 13 that these quantities are numbers that describe the theoretical average and standard deviation for data that can be modeled by the discrete random variable. Similar theoretical averages and standard deviations can be defined for continuous random variables. The precise definitions extend the formulas in Definitions 13.4 and 13.5 of Chapter 13 by replacing the summations in those formulas with suitable definite integrals. We will not go into the technical details, as they do not add a great deal to our goal of understanding how to use these concepts. However, it is important to state the results for the standard normal distribution.

Theorem 14.1: If Z is the standard normal distribution then $\mu_Z = 0$ and $\sigma_Z = 1$.■

This result has an important visual interpretation in terms of the graph of the density function $f(z)$. Referring to Figure 14.2, the mean $\mu = 0$ is clearly the center and peak point in the density graph. Notice also that the graph has two inflection points. The figure suggests, and a little calculus confirms, that these occur at the points $z = \pm 1$. Thus, the distance of the inflection points from the center is equal to the standard deviation. Moreover, since the standard deviation is one, a value such as $z = 2$ is two standard deviations above the mean zero; $z = -2.5$ is 2.5 standard deviations below the mean zero, etc. With this in mind we can give a working definition of an arbitrary normal distribution.

Definition 14.3 (Normal Distribution): A random variable X has a *normal distribution* with mean μ and standard deviation σ, if its density curve arises from the standard normal density by:

a) shifting the latter so it is centered around μ

b) contracting or expanding the standard normal density curve so it has inflection points at $\mu \pm \sigma$

c) rescaling the height of the standard normal density so that the total area beneath it remains one.■

According to this definition, a normal distribution is described by giving its mean μ and standard deviation σ. It is convenient to designate such a quantity using the notation $X = N(\mu,\sigma)$, which is an abbreviation for the statement that the random variable X has a normal distribution with mean μ and standard deviation σ. The figure below shows examples of such normal distributions with densities determined by $\mu = 5$ and $\sigma = 2, 1, 0.5, 0.25$ respectively. In all cases of course the total area under the curve is one. Notice that in order to achieve this constant area the ordinate or density rises as the central portion of the graph narrows, in accord with property c) in the definition.

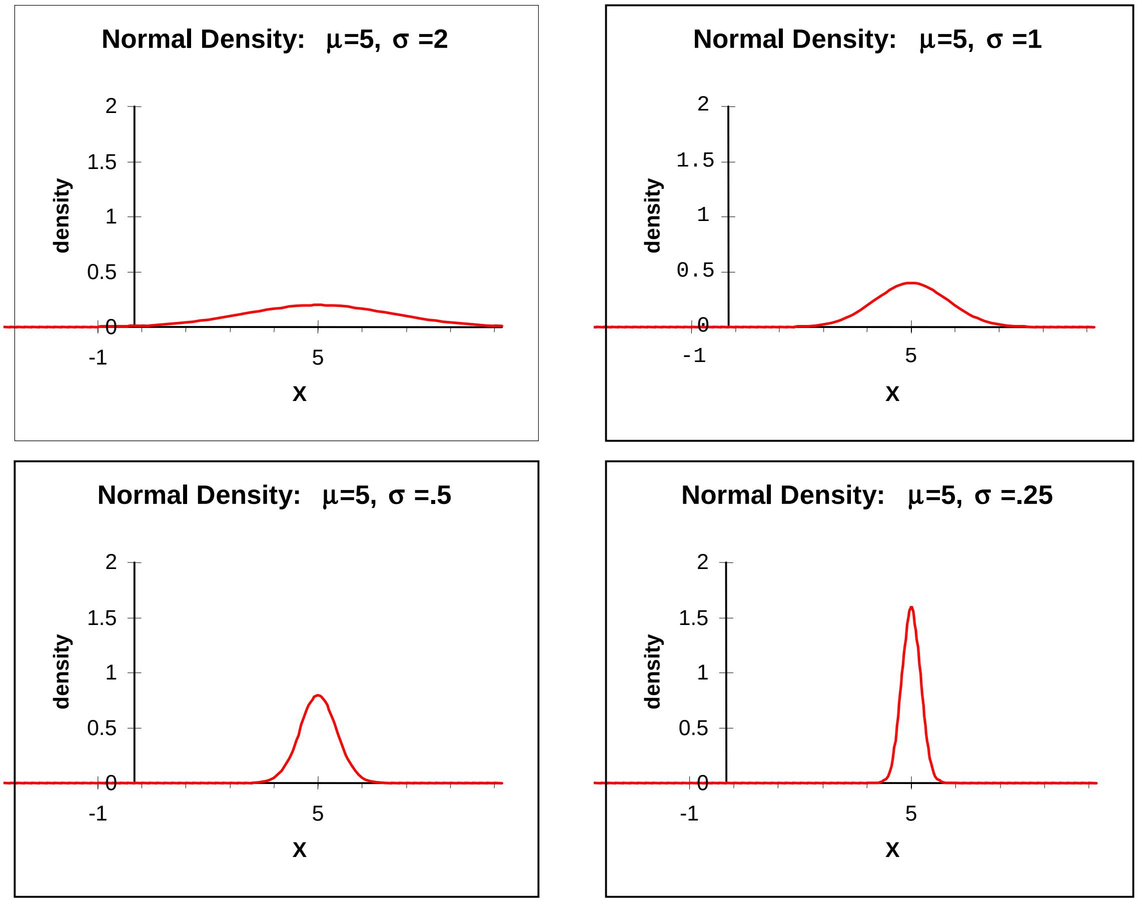


Figure 14.3

Definition 14.4 (z-score): Suppose x is a value of a random variable X that has a normal distribution with mean μ and standard deviation σ. The z*-score* of x is the number of standard deviations of x from the mean value μ. It is given by

$$z = \frac{x-\mu}{\sigma}.\blacksquare$$

Example 14.3: If $X = N(10, 3)$ what are the z-scores of $x = 12$ and $x = 7.5$?

Solution:

Using Definition 14.4, the z-score for $x = 12$ is $z = \frac{12-10}{3} \approx .67$ and the z-score for $x = 7.5$ is $z = \frac{7.5-10}{3} \approx -.83$.■

The z-score gives the number of standard deviations of a value of X from the mean. For the standard normal distribution we saw that the value of Z also equaled the number of standard deviations from the mean. The connection between these facts is the basis of the computation of probabilities for normal distributions.

Theorem 14.2: If X has a normal distribution with mean μ and standard deviation σ then the z-scores derived from X by the equation $z = \frac{x-\mu}{\sigma}$ have a standard normal distribution. Moreover for any X values a and b if z_a and z_b denote the z-scores of a and b respectively we have

$$P(a < X < b) = P(z_a < Z < z_b)$$

$$P(a < X) = P(z_a < Z)$$

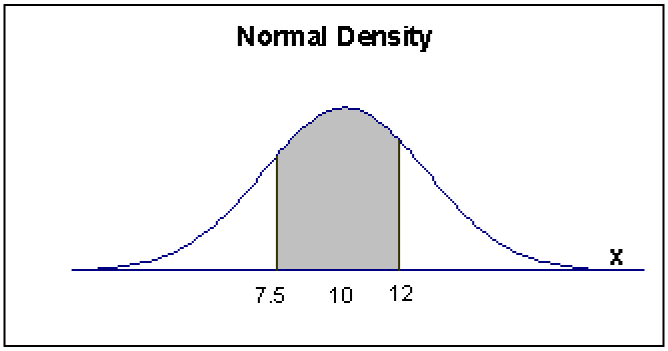
$$P(X < b) = P(Z < z_b).\blacksquare$$

The importance of this theorem cannot be overstated. In essence it says that for a normal distribution the standard deviation is the appropriate measuring stick needed to compute probabilities.

Example 14.4: For the random variable $X = N(10, 3)$ of Example 14.3 find the $P(7.5 < X < 12)$.

Solution:

Although not essential, the reader is encouraged to make a simple sketch of the bell-shaped distribution of X with the mean value clearly marked and an indication of the region whose area needs to be computed. In this example the following sketch would be sufficient and indicates we might expect a probability of approximately 0.5.



According to Theorem 14.2, to compute $P(7.5 < X < 12)$ we need the z-scores of the endpoints of the interval. These were computed in Example 14.3 as $z = -.83$ for $x = 7.5$ and $z = .67$ for $x = 12$. Therefore by Theorem 14.2 we have

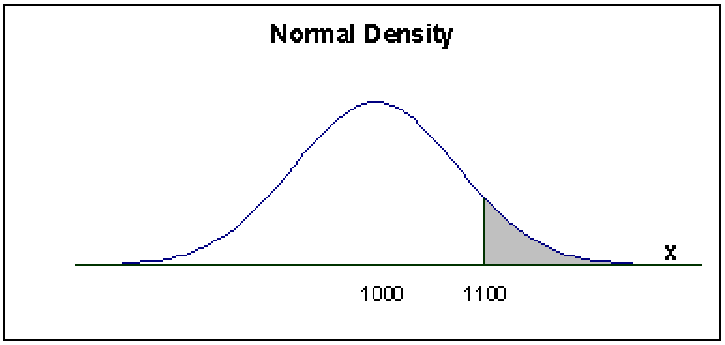
$$P(7.5 < X < 12) = P(-.83 < Z < .67) = P(Z < 0.67) - P(Z < -0.83) = .7517 - .2033 = .5484$$

which agrees fairly well with our guess based on the sketch.■

Example 14.5: The package of a light bulb states that the bulb will burn on average for 1000 hours. Suppose, in addition, that you know the standard deviation in bulb life is 75 hours and the bulb life has a normal distribution. What are the chances that a random bulb will burn for more than 1100 hours?

Solution:

We start by defining an appropriate random variable. Thus, let X denote the lifetime for an arbitrary bulb. The conditions of the problem tell us that X has a normal distribution with $\mu = 1000$ and $\sigma = 75$, in other words that $X = N(1000, 75)$. We want the probability that for an arbitrary or randomly selected bulb $X > 1100$. Before computing this probability, we make a rough sketch of the density function for X, indicating the region whose area needs to be computed. For the example at hand we have the sketch



and one might venture an estimate of 0.25 or less for the probability. To actually compute $P(X > 1100)$ we proceed as in Theorem 14.2. We need the z-score of the value $x = 1100$. This is $z = \frac{1100 - 1000}{75} \approx 1.33$. Thus by the theorem we have

$$P(X > 1100) = P(Z > 1.33) = 1 - P(Z < 1.33) = 1 - .9082 = .0918.$$

In rough terms, there is slightly less than a 10% chance that the bulb will last as long as 1100 hours.■

Theorem 14.2 provides the basis of the Bell Curve Rule (Rule 7.1).

Theorem 14.3 (Bell Curve Rule): If X has a normal distribution with mean μ and standard deviation σ then the following statements hold:

a) There is a 68% probability that X will yield a value within one standard deviation of the mean.

b) There is a 95% probability that X will yield a value within 1.96 standard deviations of the mean.

c) There is a 99.7% probability that X will yield a value within 3 standard deviations of the mean.

Proof:

We illustrate b). The proofs of the others are similar. The event that X is within 1.96 standard deviations of its mean signifies that the value of X falls in the interval $\mu - 1.96\sigma$ to $\mu + 1.96\sigma$. We therefore want $P(a \le X \le b)$ where $a = \mu - 1.96\sigma$ and $b = \mu + 1.96\sigma$. The z-scores of a and b are $\frac{a-\mu}{\sigma} = -1.96$ and $\frac{b-\mu}{\sigma} = 1.96$ respectively. Therefore,

$$P(a \le X \le b) = P(-1.96 \le Z \le 1.96) = .9750 - .0250 = .95.■$$

The precise spread of 1.96 standard deviations given in b) is often rounded, as we have done in Chapter 7, to two standard deviations when we apply the rule to data. Recall that the random variable X is a model for the description of data and the data, average and standard deviation are approximations to the theoretical values. Thus, for any data exhibiting an approximate bell-shaped histogram the Bell Curve Rule should give approximate percentages of the data lying within one, two and three standard deviations of the mean.

Example 14.6: For the normal distributions shown in Figure 14.3 estimate or use the tables to find $P(4 < X < 6)$.

Solution:

In the top left panel we have $X = N(5,2)$. The z-scores of 4 and 6 are -0.5 and 0.5 respectively. Thus $P(4 < X < 6) = P(-.5 < Z < .5) = .383$, a value obtained from the tables. In the upper right panel we have $X = N(5,1)$. The z-scores of 4 and 6 are then -1 and 1 so the Bell Curve Rule (Theorem 14.3) applies, yielding $P(4 < X < 6) = .68$. For the lower left panel, $X = N(5,0.5)$ and therefore the z-scores are -2 and 2. The Bell Curve Rule gives the estimate of .95 for the probability. Finally in the last panel, $X = N(5,.25)$ and the z-scores are -4 and 4. These z values are not given in the table, but since we are given entries for $z = \pm 3.9$ we may use these probabilities to infer that to three decimal places $P(Z < 4) = 1$ and $P(Z < -4) = 0$. This implies that $P(4 < X < 6) = 1$, as is suggested by the graph.■

Example 14.7: The scores on a test used for admissions to a selective school are approximately normally distributed with a mean of 320 and a standard deviation of 50 (The maximum possible grade is 500). What cut-off score should be set so that only the top 10% of applicants will be accepted?

Solution:

Let X be the score of a randomly selected student. The assumptions in the problem imply that we may assume $X = N(320,50)$. We want to find a value x of X so that only 10% of students score higher than this value. In other words, we want x to satisfy $P(X > x) = .10$. The z-score for the unknown value x is $z = \frac{x-320}{50}$. Using the relationship between probabilities for X and probabilities for Z we have $.10 = P(X > x) = P(Z > \frac{x-320}{50})$. However, in Example 14.2e) we showed how to find the value z_0 such that $P(Z > z_0) = .10$. We found that $z_0 = 1.28$. Thus the z-score $\frac{x-320}{50} = 1.28$. Solving for x yields the cut-off score of 384.

Having solved the problem, the solution can be explained more directly. For the Z distribution, there is only a 10% chance of a value exceeding 1.28. This is the content of Example 14.2e). For an arbitrary normal distribution X, the z-score is the number of standard deviations from the mean. Therefore, we can assert that there is a 10% chance that a value of X will be more than 1.28 standard deviations above the mean. Hence, the cut-off point must be 1.28 standard deviations above the mean 320. This gives $x = 320 + 1.28(50) = 384$.■

14.5 The Normal Approximation to the Binomial

The normal distribution was actually discovered at the end of the 18th century in the attempt to find good numerical approximations to cumulative binomial probabilities, particularly when the number of trials n was large. Although from the computational point of view the result is no longer needed if a computer is available, the result shows an important relationship between these two types of random variables. The upshot of those investigations is the following theorem.

Theorem 14.4 (Normal-Binomial Approximation I): Suppose n is large and p is neither very close to zero nor very close to one, (we'll make this precise below). The binomial distribution X with n independent trials and probability of success p can be approximated by a normal distribution with the same mean (np) and the same standard deviation $(\sqrt{npq})$ as X. More explicitly, denoting by $Y = N(np, \sqrt{npq})$ the normal distribution with mean np and standard deviation $\sqrt{npq}$, we have for any a and b that $P(a \le X \le b) \approx P(a \le Y \le b)$.■

Roughly speaking, the theorem asserts that when n is large the binomial distribution behaves like a normal distribution with mean np and standard deviation $\sqrt{npq}$.

Example 14.8: Suppose that X is a binomial distribution with $n = 100$ and $p = 0.4$. Use Theorem 14.4 to estimate $P(30 \le X \le 45)$.

Solution:

The mean value of X is $np = 40$ and the standard deviation is $\sqrt{npq} = \sqrt{100(.4)(.6)} \approx 4.9$. Theorem 14.4 asserts that the normal distribution $Y = N(40, 4.9)$ can be used to provide a good approximation to $P(30 \le X \le 45)$. The graph below, which superimposes the normal density curve on the binomial histogram, illustrates the construction.

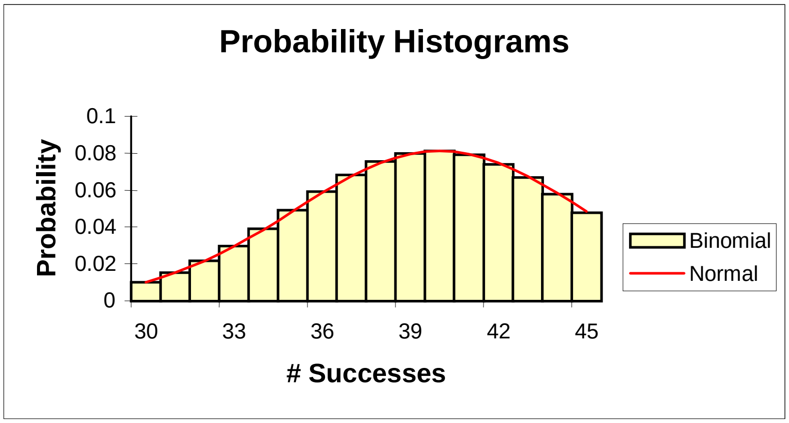


Figure 14.4

As we discussed in section 14.1, the binomial probability between 30 and 45 can be found by adding the areas of the rectangular probability bars (since these have been expanded to width one). The picture clearly shows that this area is very close to the area under the corresponding normal distribution curve. Let's perform the approximation and then, using *Excel*, compare the answer to an exact result.

According to the theorem, $P(30 \le X \le 45) \approx P(30 \le Y \le 45)$, where $Y = N(40, 4.9)$. The normal probability is computed using the methods of the previous section. We find the z-scores for the endpoints 30 and 45. These are

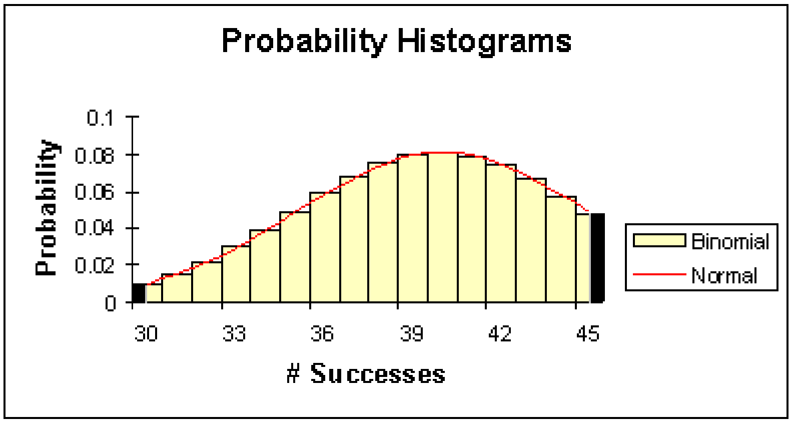
$$\frac{30-40}{4.9} = -2.04 \text{ and } \frac{45-40}{4.9} = 1.02.$$

Therefore, $P(30 \le Y \le 45) = P(-2.04 \le Z \le 1.02) = .8461 - .0207 = .8254$. The latter number is our approximation. The exact value using the *binomdist* function is .8541. The estimate differs by about .03 or 3% points - good, but not spectacular. In fact we can do better!■

Example 14.9 (The Continuity Correction): We show how to improve the approximation in the last example using a method known as the *continuity correction*.

*Solution***:**

Consider again Figure 14.4 reproduced with slight modification below.



The rectangular probability bars at the beginning and end have been split. The black portion of these bars is part of the area that is included in the probability $P(30 \le X \le 45)$, which adds the areas of all the rectangles. However, the approximation $P(30 \le Y \le 45)$ includes only the area under the normal curve from 30 to 45, and the black bars are not included in this area since they extend from the center of the box (30 or 45) to the left or right endpoints (29.5 or 40.5). Thus it would seem that our estimation of the binomial area would fall short of the actual value. This was indeed the case in Example 14.8. We can attempt to correct this by extending the interval over which we compute the area under the normal curve. Instead of going from 30 to 45 we should compute the area from the lower boundary of the left rectangle, 29.5, to the upper boundary of the right rectangle, 40.5. Thus we make the refined estimate

$$P(30 \le X \le 45) \approx P(29.5 \le Y \le 45.5) = P(-2.14 \le Z \le 1.12) = .8524,$$

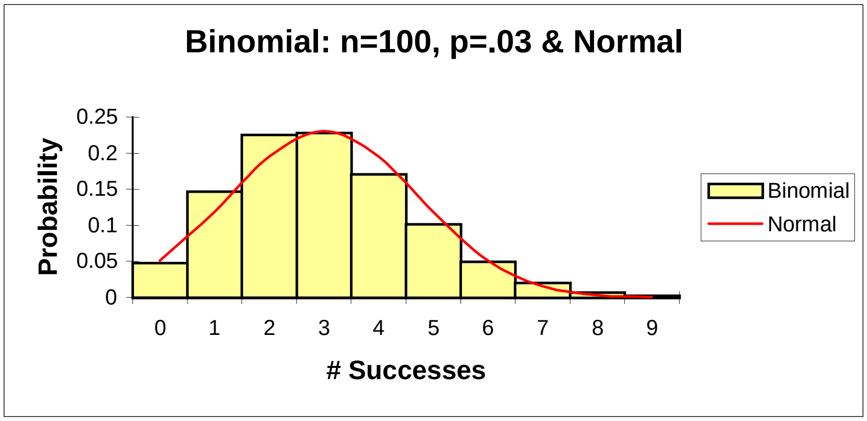
which agrees much more closely with the exact answer .8541.■

The estimation technique described in Example 14.9 is called the *continuity correction*, since it is the result of trying to fit a continuous graph to a step-like histogram. We include this in our second version of the Normal Approximation theorem. When you want to compute the binomial probabilities with greater accuracy, you should use this continuity correction. If great accuracy is not so important, use the cruder approximation given in Theorem 14.4.

Theorem 14.5 (Normal-Binomial Approximation II): If X has a normal distribution with $n > 30$ and the probability p satisfies $np > 5$ and $nq > 5$ (where $q = 1 - p$) then the normal distribution $Y = N(np, \sqrt{npq})$ can be used to approximate probabilities for X. Namely, for any a and b, the continuity correction gives

$$P(a \le X \le b) \approx P(a - 0.5 \le Y \le b + 0.5) \blacksquare$$

Theorem 14.5 clarifies the conditions under which the approximation is valid. The condition related to p, namely that both np and nq are larger than 5, asserts that the expected number of successes and expected number of failures should not be too small. The figure below illustrates what happens when $np < 5$.



The fit is very poor. The binomial distribution is quite skewed in a case such as this and a symmetric distribution such as the normal cannot adequately account for this. In this situation we have seen in Chapter 13 section 13.4 that the Poisson distribution gives an easy to compute numerical approximation to the binomial.

Example 14.10: If a coin is tossed 500 times would it be surprising to get 270 or more heads? Explain.

Solution:

The number of heads obtained has a binomial distribution with $n = 500$ and $p = .5$. According to our approximation theorems this random variable is approximately normal with mean $np = 250$ and standard deviation $\sqrt{npq} = 11.2$. An outcome exceeding 270 heads would be more than 1.8

standard deviations above the expected value. From the tables of the standard normal distribution the chance of this happening is approximately 0.04. Thus, such a result would be considered somewhat unusual.■

14.6 Tech Notes

Simulated values of the uniform distribution and various normal distributions can be obtained by using the *Data Analysis* tools and the option *Random Number Generation*, as described in Chapter 13. Random numbers in the interval [0,1] can also be generated directly in a spreadsheet using the command =*rand().*

Probabilities for normal distributions can be obtained from the file *distributions.xls*, using the *Normal* sheet. The file also contains sheets that show the comparison of the binomial and normal distributions. On any spreadsheet the commands *normsdist* and *normdist* can be used to find cumulative normal probabilities for the standard normal distribution and for an arbitrary normal distribution. The function wizard can assist you in entering the correct arguments for these functions.

Example 14.11: Use *Excel* commands to compute $P(15 \le X \le 27)$ where X is normal with $\mu = 20$ and $\sigma = 4$.

Solution:

We must compute the answer using our usual procedure of expressing the probability $P(15 \le X \le 27)$ as a difference of two cumulative probabilities. This can be accomplished by entering the commands shown below in column A, rows 1 to 3. Column B shows the output from these commands.

	A	**B**
1	=normdist(27,20,4,true)	.959941
2	=normdist(15,20,4,true)	.10565
3	=a1-a2	.854291

■

In some of our examples it has been necessary to find a z value such that $P(Z \le z)$ equaled a specified probability. We called this reverse table look-up. Mathematically this process corresponds to the computation of an inverse function. This inverse process is programmed in *Excel* through the command *normsinv*. For example, the solution of Example 14.2e) can be obtained by first converting the question to the cumulative probability statement $P(Z \le z_0) = .9$ and then entering the command =*normsinv(.9).*

14.7 Summary

Many random quantities produce numerical values that fluctuate over an interval of values, for example:

- the annual rainfall at a locale
- a randomly selected person's height
- a randomly selected person's age at death
- the site of an accident on a long stretch of highway
- time between exposure to a disease and onset of illness
- the midday CO concentration at an air quality monitoring site

These examples illustrate quantities that may be thought of as continuous random variables, X. To create a probability model for such random variables, we use a ***density function*** $f(x)$. This is a positive valued function for which the total area under the graph is one. Probabilities of the form $P(a \leq X \leq b)$ are computed using areas under the graph of the density function $f(x)$.

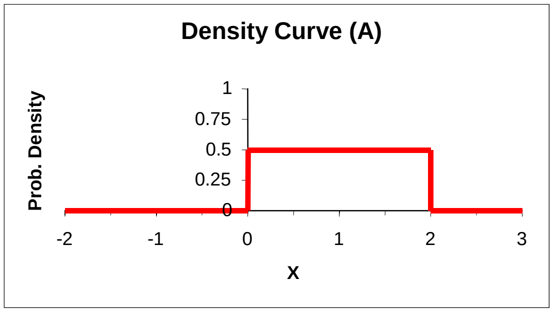
The ***uniform density*** and the ***normal density*** are two important examples of such density functions. With a uniform density, probabilities are "spread" over a certain interval in such a way that an outcome has equal chance of landing in subintervals of the same length. We say in this case that the random variable has a ***uniform distribution***. The accident site on a long stretch of more-or-less similar highway might conform to this probability model; any mile-long portion of road should be equally likely to be the site of an accident. The uniform distribution on [0,1] forms the basis of all computer simulations of random processes.

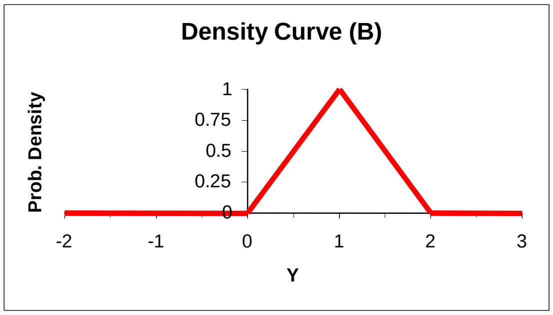
In contrast, for a normal density the outcomes cluster near the peak of the density curve, with values farther from the center being less likely to occur. The exact computation of probabilities using a normal density makes use of the ***mean*** and ***standard deviation*** of the random variable. Specifically, when *X* has a normal density curve (so we say *X* has a ***normal distribution*** or is normally distributed) we associate with each value *x* of the random variable a ***z-score***, defined as the number of standard deviations of *x* above or below its mean. These *z*-scores have probabilities described by the ***standard normal density***, with mean zero and standard deviation one, for which extensive tables or computer code can be used to find probabilities.

Normal or approximately normal distributions are common models for continuous data. Often, the applicability of a normal model is based on analogy to well-known examples that exhibit such behavior. When a large amount of data is available, the fit of the normal model can be assessed using a histogram or other graphical means. Certain discrete random variables can also be approximated using a normal model. Besides being computationally useful, such results offer important theoretical insight into the behavior of the underlying random variables. For example, a normal distribution gives good approximations to the binomial distribution when *n* is large and *p* is neither very close to zero or one.

14.8 Exercises

1. a) Using Definition 14.1 verify that the graphs in panels A and B below are each density functions for random variables, called respectively X and Y.





b) For each of the random variables described in a) find the probability that the variable takes on a value in the interval [1, 1.5].

c) Find a value x_0 such that $P(X \le x_0) = .9$

d) Find a value y_0 such that $P(Y \le y_0) = .9$

2. During a certain hour of the day, an average of 200 cars pass through a toll plaza. Assume the arrival times are spread uniformly throughout the hour and the cars are traveling to their destinations independently of each other.

a) For a fixed one-minute time period, what is the probability p that a randomly selected car will arrive at the plaza during that minute, as opposed to any other minute in the one-hour block?

b) For a fixed one-minute time period explain why the number of cars from the group of 200 arriving during that one-minute time period can be modeled by a Poisson random variable with $\lambda \approx 3.33$.

c) Dividing the hour into 60 consecutive one-minute subintervals, based on b) estimate in approximately how many of these intervals there will be 0, 1, 2, … 6, 7, 8 cars arriving.

d) The following table gives the results of a simulation of the process described in this problem (200 cars arriving at random times during a one-hour time interval). The top row specifies the number of cars arriving in a given minute and the bottom row tabulates the number of one-minute intervals with this number of cars arriving.

# of cars k arriving during one minute	0	1	2	3	4	5	6	7	8
# of one-minute intervals with k cars arriving	3	6	8	17	15	4	4	1	2

Table 14.2

Compare the frequencies in row 2 for the simulated data with the prediction based on c). Do you think the simulated data conforms to the model?

3. From Chapter 13 we know that a Poisson random variable X satisfies $\mu_X = \lambda$ and $\sigma_X = \sqrt{\lambda}$. It follows that the square of the standard deviation, the *variance* of X, satisfies $\text{var}_X = \sigma_X^2 = \lambda$. In other words, if X has a Poisson distribution then $\text{var}_X = \mu_X$. Ecologists use the ratio $\dfrac{s^2}{\overline{x}}$, the so-called *index of dispersion*, as a means of testing how well the data of an occupancy pattern conforms to a Poisson distribution. A value of this ratio that is close to one is evidence supporting a Poisson occupancy pattern.

   a) Compute the index of dispersion for the occupancy data in Table 14.1 (page 268).

   b) Compute the index of dispersion for the occupancy data in Table 14.2 above.

   c) How strongly do the indices of dispersion computed in a) and b) support the hypothesis of a Poisson distribution?

4. The number of years a person lives after receiving treatment for a type of cancer is a continuous random variable, which we denote by T. In many cases this random variable has a density function of the form

$$f(t) = \begin{cases} 0 & \text{if } t \le 0 \\ \alpha e^{-\alpha t} & \text{if } t > 0 \end{cases} \tag{14.1}$$

   Here α is a constant greater than zero and t is measured in years. We will assume that $f(t)$ satisfies Definition 14.1, and hence can be used as a density function.

   a) Sketch on the same axes the graphs of the density functions (14.1) when $\alpha = 0.5$ and when $\alpha = 0.25$.

   b) If $\alpha = 0.5$ find $P(0 \le T \le 2)$. What does the latter probability signify in terms of the patient's survival following treatment?

   c) It can be shown that the expected survival time μ_T for a random variable having the density (14.1) is $1/\alpha$. When $\alpha = .5$, what is the probability that a patient will survive for more than twice the average time? (Hint: Consider the complementary event.)

5. If Z has a standard normal distribution find

   a) $P(Z < 1.6)$

   b) $P(Z > -.52)$

   c) $P(1.2 \le Z < 2.24)$

   d) $P(|Z| < 2.5)$

6. If X has a normal distribution with mean μ =18 and $\sigma = 3$ find

   a) $P(16 \le X \le 20)$

b) $P(X \le 19)$

7. If X has a binomial distribution with $n = 50$ and $p = .3$ use the normal approximation (with continuity correction) to estimate

   a) $P(X \le 12)$

   b) $P(10 \le X \le 18)$

8. Let Z have a standard normal distribution. Find a value of z_0 such that

   a) $P(Z < z_0) \approx 0.6$

   b) $P(|Z| < z_0) \approx 0.6$

   c) If $X = N(5, 2)$ use a) to find a value x_0 such that $P(X < x_0) \approx 0.6$

9. Suppose that on average 3% of a manufacturer's output is defective. The manufacturer's distributor will return any shipments that contain more than 4.5% defectives. If shipments are made in lots of 1000, what is the probability that a shipment will be returned? (Hint: 4.5% defectives in a shipment of 1000 means 45 defectives. How likely is this, given the manufacturer's claim?)

10. Suppose the annual rainfall in N.Y. is normally distributed with a mean of 42 inches per year and a standard deviation of 4 inches. What is the probability that in a given year the rainfall, X, is between 36 and 48 inches?

11. Suppose it is known that only 90% of the people who reserve a seat on a flight actually show up. If a plane holds 220 passengers, explain, by using a suitable binomial distribution, why the airline is fairly confident that they can reserve 235 seats and still provide a seat for everyone who shows.

12. a) The manufacturer of a TV picture tube knows that with customary usage the average lifetime for a tube is 5.4 years, with a standard deviation of 1.1 years. Assuming the lifetime is normally distributed find the percentage of tubes that will fail before 5 years.

    b) If the manufacturer wants to give a full refund warranty for defective tubes, how long should the warranty period be so that at most 0.4% of the tubes are returned under warranty?

13. Suppose the yearly income of workers in a certain industry is normally distributed with mean equal to \$30,000 and standard deviation \$4250.

    a) If a worker is chosen at random, what is the probability that the worker's income will be between \$25,000 and \$40,000?

    b) What is the probability that the worker's income will exceed \$25,000?

c) Find the income level I so that only 1% of workers will have incomes exceeding I.

14. Assume that the time required to complete a certain medical procedure is normally distributed with mean 3 hours and standard deviation of 20 minutes.

    a) What is the probability that the procedure will be completed in less than 2.5 hours?

    b) Assume the procedure is done on an outpatient basis. A friend accompanies the patient, but wishes to leave the facility while the procedure is in progress in order to complete some errands. However, the friend wants to be 98% certain she will be back before the procedure is done. What is the maximum length of time she should be gone?

15. Good-n-Tasty candy bars have a weight of 6 oz. listed on the package. In practice, the packaged bars are not individually weighed, so the 6-oz. weight listed on the package may be incorrect. In particular, the bar may actually weigh less than 6 oz. Suppose the actual weights are normally distributed with a mean of 6.25 oz. and a standard deviation of 0.1 oz.

    a) What fraction of the bars will actually have a weight below the 6-oz. weight listed on the package?

    b) In order to comply with anti-fraud statutes, the manufacturer must make sure that no more than 1 in 1000 bars has an actual weight below 6 oz.. Assuming that the standard deviation of the production process remains the same (0.1 oz.), what is the smallest value for the mean weight that will meet the anti-fraud requirements?

16. In 1977 a very large study (Hypertension Detection and Follow-up Program) produced data showing that the diastolic blood pressure reading in adults has a normal distribution with a mean of 85 mm Hg and a standard deviation of about 13. (The units refer to the height in a mercury (Hg) column typically used to measure pressure.)

    a) What fraction of the population would be classified as mildly hypertensive (i.e. having high blood pressure) if the cutoff for this category was a diastolic pressure in excess of 95 mm Hg?

    b) What fraction of the population would be classified as having severe hypertension if the cutoff for this category was a diastolic pressure in excess of 115 mm Hg?

17. The diastolic blood pressure of an individual varies from one reading to the next, even when the person is resting. Suppose a person actually has moderate hypertension, meaning his average diastolic pressure is 105 mm Hg. If studies have shown that pressure readings for individuals show a standard deviation of 7 mm Hg, what is the probability that a single reading for this individual will yield a pressure reading below 95, and hence a misclassification of the individual as only mildly hypertensive?

18. a) Use *Excel's* random number generator to obtain 1000 random values drawn from the standard normal distribution.

    b) Find the mean and standard deviation of the numbers you obtained in a). These numbers should be close to zero and one respectively.

c) Test the conclusion of the Bell Curve Rule (Theorem 14.3) for the data generated in a).

19. a) Use *Exce*l to generate 1000 random numbers X from a normal distribution $N(1,1)$. Place these in column A of a worksheet. Generate another 1000 random numbers Y from a distribution that is $N(2,1)$ and place these in column B of the same sheet.

b) In column C compute the sum $X+Y$ of the numbers generated in columns A and B above. Find the mean and standard deviation of these numbers and then using the Bell Curve Rule test whether the sums appear to conform to a normal distribution.

c) In column D compute the product XY of the random numbers obtained in columns A and B. Repeat the instructions in b) for the numbers in column D.

20. If X is a random variable taking on positive values only and $\log X$ (natural logarithm) has a normal distribution, the random variable X is said to have a *lognormal* distribution. The lognormal distribution arises in finance and risk management, as well as in ecological applications involving the dispersion of hazardous wastes and other pollutants.

a) Use *Excel's* random number generator to simulate 1000 values from the standard normal distribution. Using the latter numbers obtain random numbers X whose natural logarithm has the standard normal distribution.

b) Make a histogram of the lognormally distributed random numbers X obtained in a), using a small bin width (for example 0.5).

c) Is the histogram in a) symmetric? Is it skewed? If so, which way? Find the median and the mean of the lognormally distributed numbers X. Is the relationship between the mean and median reflective of the skewness of the distribution?