CONTENT OVERVIEW

Examples of random phenomena abound. Walk up to students at your school or campus and ask them to rate their happiness: Unhappy, So-So, Happy. Show up at Pete and Gerry's farm, randomly select a hen and weigh it. As these two examples illustrate, sometimes the outcomes of a random phenomenon are categories and other times numbers. Next, we create random variables by mapping the outcomes of random phenomena to numbers.

Start with the example of the Happiness Survey that was actually given to residents of Somerville, Massachusetts. (Refer back to Unit 13, Two-Way Tables.) We can create a random variable x as follows:

$$x = \begin{cases} 0, & \text{if Unhappy} \\ 1, & \text{if So-So} \\ 2, & \text{if Happy} \end{cases}$$

Using the data from Somerville's Happiness Survey, we can assign a probability, *p(x)*, to each of the numeric outcomes of *x*. The numeric outcomes of the random variable *x* together with their probability assignments form the probability distribution of *x* shown below.

x	0	1	2
$p(x)$	0.048	0.369	0.583

Table 20.3. Probability distribution for x.

The probabilities in Table 20.3 satisfy two properties required of all probability distributions:

$$0 \le p(x) \le 1$$

$$\sum p(x) = 1$$

In other words, the values for *p(x)*, 0.048, 0.369, and 0.583, are all between 0 and 1 and they sum to 1: 0.048 + 0.369 + 0.583 = 1.

We can represent the probability distribution *p(x)* graphically with the probability histogram in Figure 20.2. The possible values for *x* are on the horizontal axis and probability is on the vertical axis.

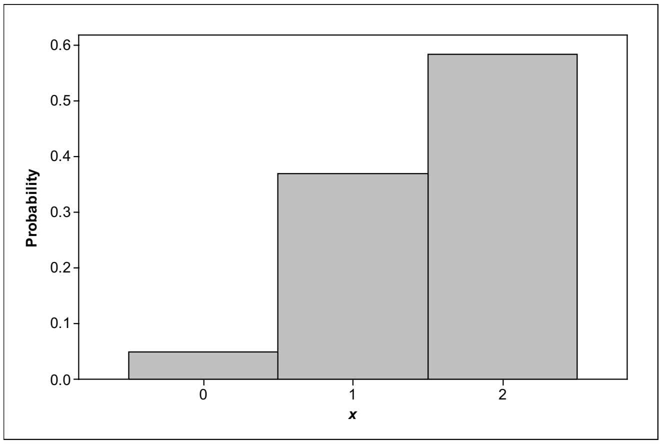

Figure 20.2. Probability histogram representing p(x).

Next, consider the random phenomenon of the weight of 7½-week old hens. In this case, the outcome is already a number and so, we can define a new random variable *w* = hen weight. Notice that *w* takes on values in the interval from the weight of the smallest hen to the weight of the largest hen. An interval contains too many numbers to list them all – so, we can't assign probabilities to each possible weight as we did with *x* in Table 20.3. From the histogram of data on hen weights (Figure 20.3), we find that a normal density curve is a good approximation for the distribution of *w*.

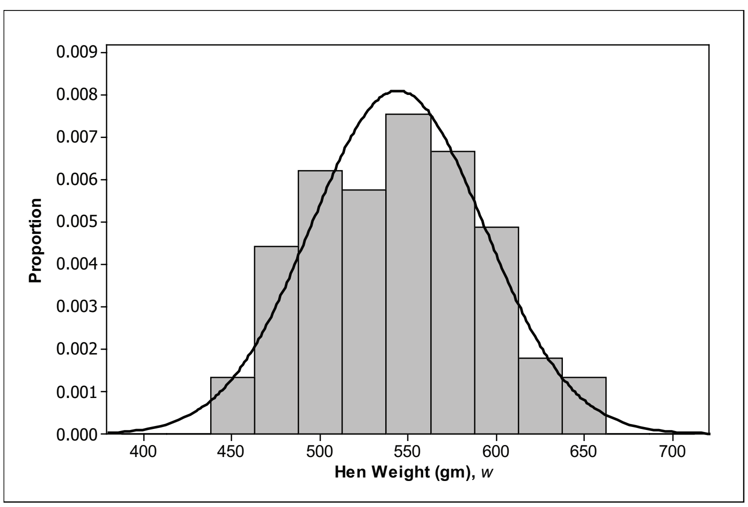

Figure 20.3. Approximating the distribution of w with a normal density curve.

The two random variables that we have looked at – *x*, happiness rating, and *w*, hen weight – are examples of two different types of random variables. Since it is possible to list all possible outcomes for *x*, *x* is called a discrete random variable. However, *w* takes on values in an interval – there are too many possible outcomes to list them all; *w* is an example of a continuous random variable.

Let's take a look at another discrete random variable. Table 20.4 gives a probability distribution for family size, *y*, in the United States. (Although there are some families that are bigger than 8, the likelihood is so small that we ignored them in this probability distribution model.)

y	1	2	3	4	5	6	7	8
$p(y)$	0.15	0.23	0.19	0.23	0.12	0.05	0.02	0.01

Table 20.4. Distribution of U.S. family size.

From the probability distribution for *y*, we can find the probability that a randomly selected family will consist of at least two people using the Complement Rule:

$$P(y \geq 2) = 1 - P(y < 2) = 1 - p(1) = 0.85$$

We can use the Addition Rule to find the probability that a randomly selected family will have two to four members:

$$P(2 \leq y \leq 4) = P(y = 2 \text{ or } y = 3 \text{ or } y = 4) = p(2) + p(3) + p(4) = 0.65$$

We can do more with probability distributions of discrete random variables than just compute probabilities. We can calculate the random variable's mean and standard deviation. All that's needed are the following formulas:

$$\mu = \sum x \cdot p(x)$$

$$\sigma^2 = \sum (x - \mu)^2 \cdot p(x) \quad \text{and} \quad \sigma = \sqrt{\sigma^2}$$

First, we calculate the mean family size. For the calculation, multiply each outcome by its probability and then sum the results:

$$\mu = 1 \times 0.15 + 2 \times 0.23 + 3 \times 0.19 + 4 \times 0.23 + 5 \times 0.12 + 6 \times 0.05 + 7 \times 0.02 + 8 \times 0.01$$

$$\mu = 3.22$$

Next, we calculate the variance. For this calculation, multiply each outcome's squared deviation from the mean by its probability and then sum the results:

$$\sigma^2 = (1 - 3.22)^2(0.15) + (2 - 3.22)^2(0.23) + \ldots + (8 - 3.22)^2(0.01)$$

$$\sigma^2 \approx 2.512$$

To get the standard deviation, take the square root of the variance:

$$\sigma \approx \sqrt{2.512} \approx 1.58$$

Now, we return to the problem of determining probabilities for a continuous random variable, such as hen weight. For that we need the probability density curve. We will assume hen weight is normally distributed with mean 544 grams and standard deviation 49 grams. We can find probabilities by calculating areas under the density curve – for these calculations we need technology or must convert to *z*-scores and use the standard normal table. For example, suppose that we want to find the probability that a randomly selected hen weighs between 500 and 600 grams, $P(500 \le w \le 600)$. Figure 20.4 shows the area under the normal density curve over the interval from 500 to 600. That area gives the probability we are seeking.

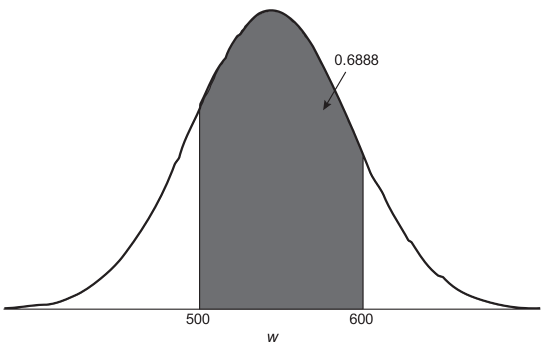

Figure 20.4. Shaded area under normal density curve.

We can use software to give us this area (as shown on Figure 20.4) or we can convert the endpoints of the interval into *z*-scores and use the standard normal table to get the probability. (This method was introduced in Unit 8, Normal Calculations.)

$z = \dfrac{500 - 544}{49} \approx -0.90$ and $z = \dfrac{600 - 544}{49} \approx 1.14$

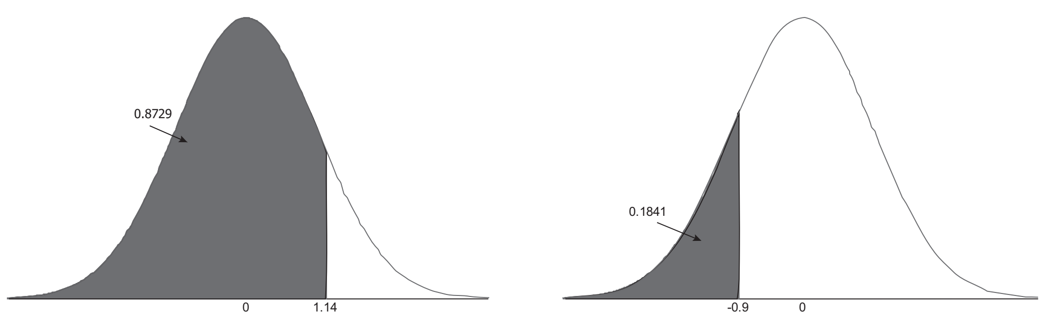

Figure 20.5. Areas to the left of z = 1.14 (a) and z = -0.90 (b).

We determine areas under the standard normal curve to the left of our two *z*-values. These areas are 0.8729 (a) and 0.0841 (b), corresponding to *z* = 1.14 and *z* = -0.90, respectively. By subtracting these two areas, we get the area under the standard normal curve over the interval from -0.90 and 1.14:

$$0.8729 - 0.1841 = 0.6888.$$

This gives us the same value that we obtained using software in Figure 20.4.

Statisticians can also compute the mean and variance of a continuous random variable. All that's needed is a formula for the probability density curve and some calculus. The need for calculus puts computing the mean and variance of continuous random variables outside of the scope of this course.