Supplement to Stat 1100 with Johnson and Wichern

Nancy Pfenning

More on Chapter 6: Discrete Random Variables

We may specify the **probability distribution** $f(x) = P(X = x)$ of a discrete random variable X as the list of its possible values x_i and their associated probilities $f(x_i)$. This list can be used to find the probability of *specific values* of X:

Example

Household size in the U.S. has the following distribution:

x	$f(x)$
1	.251
2	.321
3	.171
4	.154
5	.067
6	.022
7	.014

The probability that a household has 3 people is $P(X = 3) = f(3) = .171$.

Often, we are interested in the probability that a random variable takes a value *over an interval.* We define the **cumulative distribution function** $F(x)$ as $P(X \leq x)$. On the probability density histogram, whereas $f(x) = P(X = x)$ is represented by the area of the *single* histogram bar over x, $F(x) = P(X \leq x)$ is represented by the total area of *all* the bars to the left of (and including) the bar over x.

Example

Specify the cumulative distribution function $F(x)$ for household size in the U.S. Sketch a probability histogram for $f(x)$, and solve for the given probabilities, first using $f(x)$, then using $F(x)$.

1. Find the probability of no more than 3 people.
2. Find the probability of fewer than 3 people.
3. Find the probability of more than 4 people.
4. Find the probability of between (and including) 2 and 4 people.

The probability histogram has a bar of height .251 over $x = 1$, a bar of height .321 over $x = 2$, etc.

Next, to solve for F of any x, we sum the values of $f(x)$ up to and including the given x:

x	$F(x)$
1	.251
2	.572
3	.743
4	.897
5	.964
6	.986
7	1.000

Now we solve for probabilities, using f and then F:

1. the probability of no more than 3 people:
   $P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = f(1) + f(2) + f(3) = .251 + .321 + .171 = .743$
   alternatively, $P(X \leq 3) = F(3) = .743$.
2. the probability of fewer than 3 people:
   $P(X < 3) = f(1) + f(2) = .251 + .321 = .572$
   alternatively, $P(X < 3) = P(X \leq 2) = F(2) = .572$.
   [Note that $P(X \leq 3)$ and $P(X < 3)$ are not the same.]
3. the probability of more than 4 people:
   $P(X > 4) = f(5) + f(6) + f(7) = .067 + .022 + .014 = .103$.
   alternatively, $P(X > 4) = 1 - P(X \leq 4) = 1 - F(4) = 1 - .897 = .103$.
4. the probability of between (and including) 2 and 4 people:
   $P(2 \leq X \leq 4) = f(2) + f(3) + f(4) = .321 + .171 + .154 = .646$
   alternatively, $P(2 \leq X \leq 4) = P(X \leq 4) - P(X \leq 1) = F(4) - F(1) = .897 - .251 = .646$

Chapter 7: Continuous Random Variables and Sampling Distributions

In Chapter 2, we introduced the concept of a *density curve* as an idealization of a density histogram. Now, a **probability density curve** for a continuous random variable is a smoothed-out version of a probability histogram. Discrete random variables, such as household size, have distinct possible values x_i that can be specified in a list, along with their probabilities $f(x_i) = P(X = x_i)$. They can be displayed with a probability histogram.

Continuous random variables, such as height, have infinitely many possible values over an entire interval. The distribution of a continuous random variable is given by its **probability density function** $f(x)$, a smooth curve with the following properties:

1. The total area under the probability density curve is 1.

2. $P(a \leq X \leq b)$ =area under the probability density curve between a and b.

3. $f(x) \geq 0$ for all x (that is, density curves may not dip below the x axis).

Notice how these parallel the properties of the probability distribution function $f(x)$ for discrete random variables, as stated at the end of Section 6.3. In many respects, continous random variables are analogous to discrete ones. One important difference, however, is that, because a continuous random variable has an infinite number of point values, the probability of any particular one of them occurring is actually zero! Thus, it only makes sense to talk about the probability of a continuous random variable taking a value *over an interval.* For discrete random variables, we had to be careful about whether there was strict inequality or not when solving for probabilities over intervals. [For X = household size, we found $P(X \leq 3) = .743$, whereas $P(X < 3) = .572$.] For continuous random variables, we can afford to be careless about strict inequalities, since

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

for a continuous random variable X. The area under the curve between a and b remains the same, whether or not the endpoints are included.

Features of a Continuous Distribution

Just as with any distribution, we summarize by focusing on its *center*, *spread*, and *shape*.

To describe **center** of a continuous distribution, we may use **median** (equal area point) or **mean** (balance point). Since the total area under a density curve must be 1, the **median** is a value of X which divides the area under the curve into .5 and .5 on either side. In physics, if a flat solid has a profile described by a density function f, then the center of gravity—or balance point—is found by taking the integral $\int_{-\infty}^{+\infty} xf(x)dx$. In statistics, the *same* formula tells us the expected value (**mean**) of a continuous random variable X with probability density function $f(x)$!

$$E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx$$

Median and mean will match when the density curve is symmetric. Just as we saw for data sets in Chapter 2, mean is less than median for a left-skewed distribution and greater than median for a right-skewed distribution.

If median is used as the measure of center, we can summarize **spread** with **quartiles**, which further divide the area under the curve into fourths, and the accompanying **interquartile range** IR, which tells the range of the middle half of the distribution. If mean is our measure of center, then the accompanying measure of spread is **standard deviation**. Again, there is a physical analogy. If a flat solid has a profile described by density $f(x)$, the **moment of inertia**, or amount of energy required to stop or start the body spinning about its center of gravity μ, is found by taking the integral $\int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx$. This measures how spread out the body is around its center. Similarly, in statistics we measure the spread of a continuous distribution whose density is $f(x)$ with its **variance**

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$$

$$= \left(\int_{-\infty}^{+\infty} x^2 f(x)dx\right) - \mu^2 \qquad \text{[Shortcut formula]}$$

Standard deviation is simply

$$\sigma = \sqrt{Var(X)}$$

Shape may be symmetric or skewed (right or left). It may be bell-shaped with tapering ends, uniform, bimodal, etc. We will encounter various shapes in examples to come.

Chapter 7 Supplement A

Whether our random variable X has a discrete or continuous distribution, we are generally interested in probabilities of X taking certain values; in the mean μ (center) of X; and in the standard deviation σ (spread) of X. To find probabilities, we may work with the probability distribution/density function $f(x)$ or with the cumulative distribution $F(x)$. Formulas for μ and σ involve $f(x)$.

Problems for discrete random variables tend to be more straightforward, so we will first work through a detailed example in the discrete case, then state the basic formulas for a discrete random variable X. Next we will present the analogous formulas for a continuous random variable X, and work a detailed example in the continuous case.

Example

Let X be the roll of a single die.

1. Specify $f(x) = P(X = x)$ with a list and then a formula; graph its probability histogram.
2. Find the following probabilities:
   (a) $P(3 \leq X \leq 5)$
   (b) $P(X = 3)$

(c) $P(X \leq 3)$
(d) $P(X < 3)$
(e) $P(X > 3)$
(f) $P(X \geq 3)$

3. Verify that $\sum_{\text{all } x} f(x) = 1$.
4. Specify $F(x) = P(X \leq x)$ with a list and then a formula; give an interpretation of F with respect to the probability histogram. Re-solve the probabilities in part (2), this time using F.
5. Find $E(X) = \mu$.
6. Find $Var(X) = \sigma^2$.
7. Find σ.

Here are the solutions:

1. The probability distribution function f is easily found:

x	$f(x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

We see that $f(x) = \frac{1}{6}$ for all integers x between 1 and 6. [$f(x)$ is zero otherwise.] Its probability histogram is **uniform**: 6 bars of height $\frac{1}{6}$ over the 6 values 1, 2, 3, 4, 5, 6.

2. We can use $f(x)$ to solve for the stated probabilities:

(a) $P(3 \leq X \leq 5) = f(3) + f(4) + f(5) = \frac{3}{6}$.
(b) $P(X = 3) = f(3) = \frac{1}{6}$.
(c) $P(X \leq 3) = f(1) + f(2) + f(3) = \frac{3}{6}$
(d) $P(X < 3) = f(1) + f(2) = \frac{2}{6}$
(e) $P(X > 3) = f(4) + f(5) + f(6) = \frac{3}{6}$
(f) $P(X \geq 3) = f(3) + f(4) + f(5) + f(6) = \frac{4}{6}$

3. $\sum_{\text{all } x} f(x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$.

4. The cumulative distribution function F is found to be

x	$F(x)$
1	$\frac{1}{6}$
2	$\frac{2}{6}$
3	$\frac{3}{6}$
4	$\frac{4}{6}$
5	$\frac{5}{6}$
6	$\frac{6}{6}$

We see that $F(x) = \frac{x}{6}$ for all integers x between 1 and 6. [$F(x)$ is zero for any $x < 1$, $F(x)$ is one for any $x \geq 6$.] $F(x)$ is represented by the area of the histogram bars to the left of (and including) x.

(a) $P(3 \leq X \leq 5) = P(X \leq 5) - P(X \leq 2) = F(5) - F(2) = \frac{5}{6} - \frac{2}{6} = \frac{3}{6}$

(b) $P(X = 3) = P(X \leq 3) - P(X \leq 2) = F(3) - F(2) = \frac{3}{6} - \frac{2}{6} = \frac{1}{6}$ [Note: this problem was much easier to solve using f!]

(c) $P(X \leq 3) = F(3) = \frac{3}{6}$ [This, on the other hand, is easier to solve using F.]

(d) $P(X < 3) = P(X \leq 2) = F(2) = \frac{2}{6}$

(e) $P(X > 3) = 1 - P(X \leq 3) = 1 - F(3) = 1 - \frac{3}{6} = \frac{3}{6}$

(f) $P(X \geq 3) = 1 - P(X \leq 2) = 1 - F(2) = 1 - \frac{2}{6} = \frac{4}{6}$

5. $E(X) = \mu = \sum_{\text{all } x} xf(x) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$
6. $Var(X) = \sigma^2 = \sum_{\text{all } x} x^2 f(x) - \mu^2$

$$= 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) - 3.5^2 = 2.9$$

7. $\sigma = \sqrt{2.9} = 1.7$

Thus, the number rolled on a die is about 3.5, give or take about 1.7.

Note: The cumulative distribution function F may at this point seem to be more trouble than it's worth. This may be true in the discrete case, but in the continuous case F is found by integrating, rather than by summing. Once we have an expression for F, we are spared the work of integrating the density f every time we want to solve for a probability.

The following formulas summarize our results for discrete random variables from Chapter 6.

Formulas for Discrete Random Variables

1. Graph: the height of the probability histogram is $f(x) = P(X = x)$ (the probability distribution).
2. Probabilities: $P(a \leq X \leq b) = \sum_{x=a}^{b} f(x)$.
3. Law: $\sum_{\text{all } x} f(x) = 1$.
4. Cumulative Distribution Function: $F(x) = P(X \leq x)$ so $P(X \leq a) = F(a) = \sum_{\text{all } x \leq a} f(x)$.

   $F(x) = 0$ for x less than the lowest possible value of X and $F(x) = 1$ for x greater than the highest possible value.
5. Mean: $E(X) = \mu = \sum_{\text{all } x} xf(x)$.
6. Variance: $Var(X) = \sigma^2 = \sum_{\text{all } x}(x - \mu)^2 f(x) = \left(\sum_{\text{all } x} x^2 f(x)\right) - \mu^2$
7. Standard Deviation: $\sigma = \sqrt{Var(X)}$

Note the similarities to the following:

Formulas for Continuous Random Variables

1. Graph: the height of the probability density curve is $f(x)$ (the probability density function).
2. Probabilities: $P(a \leq X \leq b) = \int_{x=a}^{b} f(x)dx$.
3. Law: $\int_{-\infty}^{+\infty} f(x)dx = 1$.
4. Cumulative Distribution Function: $F(x) = P(X \leq x)$ so $P(X \leq a) = F(a) = \int_{-\infty}^{a} f(x)dx$

   for any specified value a; in general, $F(x) = \int_{-\infty}^{x} f(t)dt$ for any x.

   $F(x) = 0$ for x less than the lowest possible value of X and $F(x) = 1$ for x greater than or equal to the highest possible value.

5. Mean: $E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx$.
6. Variance: $Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \left(\int_{-\infty}^{+\infty} x^2 f(x)dx\right) - \mu^2$
7. Standard Deviation: $\sigma = \sqrt{Var(X)}$

Observe that the continuous formulas parallel their discrete counterparts. In the discrete case, we solve for probabilities by summing areas of rectangles whose heights are given by f, whereas in the continuous case, we solve for probabilities by finding the area underneath the density function f, that is, by integrating f. The mean of X is a weighted average of its values: sum up $xf(x)$ for all possible x in the discrete case; integrate $xf(x)$ over the range of possible values of x in the continuous case. [This is equivalent to integrating $xf(x)$ from minus to plus infinity, since $f(x)$ is taken to be zero outside the range of possible values of x.] Formulas for variance and standard deviation are also comparable from the discrete to the continuous case. In both cases, we will opt for the shortcut formula in practice.

Example

Let X be the point on a 5-mile stretch of highway where the next accident occurs, and assume that all locations are equally likely. Then $f(x) = \frac{1}{5}$ for all x between 0 and 5. [$f(x) = 0$ otherwise.]

1. Graph the probability density function $f(x)$.
2. Find the following probabilities:
   (a) $P(3 \le X \le 4)$
   (b) $P(X = 3)$
   (c) $P(X \le 3)$
   (d) $P(X < 3)$
   (e) $P(X > 3)$
3. Verify that $\int_{-\infty}^{+\infty} f(x)dx = 1$.
4. Give an interpretation of $F(x) = P(X \le x)$ and find its formula. Use F to re-solve the probabilities in part (2).
5. Find $E(X) = \mu$.
6. Find $Var(X) = \sigma^2$.
7. Find σ.

Here are the solutions:

1. The graph of f is **uniform**: all values of x over the interval from 0 to 5 are equally likely, so the height of f is constant at $\frac{1}{5}$.
2. (a) $P(3 \le X \le 4) = \int_3^4 \frac{1}{5}dx = \frac{x}{5}\Big|_3^4 = \frac{4}{5} - \frac{3}{5} = \frac{1}{5}$
   (b) $P(X = 3) = 0$ [The area under the curve over any single point is zero.]
   (c) $P(X \le 3) = \int_{-\infty}^{3} f(x)dx = \int_{-\infty}^{0} 0dx + \int_0^3 \frac{1}{5}dx = 0 + \frac{x}{5}\Big|_0^3 = \frac{3}{5} - \frac{0}{5} = \frac{3}{5}$.
   Note: Our density function f is often defined on a prescribed interval, and is zero outside that interval. Since integrating zero over any interval results in zero, we often save ourselves the trouble of writing out an integral all the way down to $-\infty$ or all the way up to $+\infty$. Instead, we simply write our lower limit of integration as the smallest value of X for which f is non-zero, and/or the upper limit as the largest value of X for which f is non-zero.
   (d) $P(X < 3) = P(X \le 3) = \frac{3}{5}$ [For continuous random variables X, strict inequality or not doesn't affect probabilities.]

(e) $P(X > 3) = \int_3^5 \frac{1}{5} dx = \frac{x}{5}\Big|_3^5 = \frac{5}{5} - \frac{3}{5} = \frac{2}{5}$

3. $\int_{-\infty}^{+\infty} f(x)dx = \int_0^5 \frac{1}{5} dx = \frac{x}{5}\Big|_0^5 = \frac{5}{5} - \frac{0}{5} = 1.$
4. $F(x) = P(X \leq x)$ is the area under the curve $f(x)$ up to the point x. For x between 0 and 5,

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_0^x \frac{1}{5} dt = \frac{t}{5}\Bigg|_0^x = \frac{x}{5} - \frac{0}{5} = \frac{x}{5}.$$

$F(x)$ is zero for any x less than 0, and one for any x greater than or equal to 5.

(a) $P(3 \leq X \leq 4) = P(X \leq 4) - P(X \leq 3) = F(4) - F(3) = \frac{4}{5} - \frac{3}{5} = \frac{1}{5}$

(b) $P(X = 3) = 0$

(c) $P(X \leq 3) = F(3) = \frac{3}{5}$

(d) $P(X < 3) = P(X \leq 3) = F(3) = \frac{3}{5}$

(e) $P(X > 3) = 1 - P(X \leq 3) = 1 - F(3) = 1 - \frac{3}{5} = \frac{2}{5}$.

5. Mean: $E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^5 x(\frac{1}{5})dx = \frac{1}{10}x^2\Big|_0^5 = \frac{25}{10} - 0 = 2.5.$
Because a uniform distribution is symmetric, 2.5 is also the median, or equal-area point.
6. Variance:

$$Var(X) = \sigma^2 = \left(\int_{-\infty}^{+\infty} x^2 f(x)dx\right) - \mu^2$$

$$= \int_0^5 \frac{x^2}{5} dx - 2.5^2 = \frac{1}{15}x^3\Bigg|_0^5 - 6.25 = \frac{125}{15} - 6.25 = 2.08$$

7. Standard Deviation: $\sigma = \sqrt{Var(X)} = \sqrt{2.08} = 1.44$. This is the "typical" distance of values of X from their mean 2.5.

In the discrete case, $f(x)$ is the height of the probability histogram bars and $F(x)$ gives probabilities by summing up areas of the bars. In the continuous case, $f(x)$ is the height of the probability density curve and $F(x)$ gives probabilities as areas under the curve—it is the integral of f. Conversely, f is the derivative of F!

Example

Show how f and F from our previous example are related. For x between 0 and 5,

$$\text{If} \quad f(x) = \frac{1}{5} \quad \text{then} \quad F(x) = \int_{-\infty}^{x} f(t)dt = \int_0^x \frac{1}{5}dt = \frac{t}{5}\Bigg|_0^x = \frac{1}{5}x$$

$$\text{If} \quad F(x) = \frac{1}{5}x \quad \text{then} \quad f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}\frac{1}{5}x = \frac{1}{5}$$

Supplemental Exercises A

Before doing the exercises below, students should do 7.2 and 7.4 on page 333 of the textbook.

1. Refer to the probability distribution in Exercise 6.83 on page 321 of your text.

   (a) Solve for the following probabilities, using the probability distribution f:

      i. $P(X < 2)$

      ii. $P(X \leq 2)$

      iii. $P(X > 1)$

      iv. $P(0 < X \leq 2)$

   (b) Make a table with values of the cumulative distribution function $F(x)$.

(c) Re-solve the probabilities in part (a), now using F.

2. Refer to the probability distribution in Exercise 6.84 on page 321 of your text.
   (a) Solve for the following probabilities, using the probability distribution f:
      i. $P(X < 4)$
      ii. $P(X \leq 4)$
      iii. $P(X > 4)$
      iv. $P(2 < X \leq 4)$
   (b) Make a table with values of the cumulative distribution function $F(x)$.
   (c) Re-solve the probabilities in part (a), now using F.

3. Imagine the probability density curve $f(x)$ for the random variable X = age of people who died this year in the U.S. (A sketch might help.)
   (a) Would the shape of f be left-skewed, fairly symmetric, or right-skewed?
   (b) Which is larger, mean age at death or median age at death (or should they be roughly the same)?

4. Let X be the score on an exam where most students did quite well, but a few students did unusually poorly, and imagine the probability density curve $f(x)$ for this random variable.
   (a) Would the shape of f be left-skewed, fairly symmetric, or right-skewed?
   (b) Which is larger, mean score or median score (or should they be roughly the same)?

5. Imagine the probability density curve $f(x)$ for the random variable X = age of people who died of *unnatural causes* this year in the U.S. (A sketch might help.)
   (a) Would the shape of f be left-skewed, fairly symmetric, or right-skewed?
   (b) Which is larger, mean age at death or median age at death (or should they be roughly the same)?

6. Imagine the probability density curve $f(x)$ for the random variable X = earnings last year of all graduates of Pitt's business school.
   (a) Would the shape of f be left-skewed, fairly symmetric, or right-skewed?
   (b) Which is larger, mean earnings or median earnings (or should they be roughly the same)?

7. Imagine the probability density curve $f(x)$ for the random variable X = trunk length of adult female African elephants.
   (a) Would the shape of f be left-skewed, fairly symmetric, or right-skewed?
   (b) Which is larger, mean trunk length or median trunk length (or should they be roughly the same)?
   (c) What is the probability that a trunk will be exactly 4.41 feet long?

8. Imagine the probability density curve $f(x)$ for the random variable X = circumference of a sequoia tree in southern California.
   (a) Would the shape of f be left-skewed, fairly symmetric, or right-skewed?
   (b) Which is larger, mean circumference or median circumference (or should they be roughly the same)?
   (c) What is the probability that a tree will measure precisely 30.027 feet in circumference?

9. Which random variable would have a larger standard deviation—age of a randomly chosen Pitt student or age of a randomly chosen American?

10. Which random variable would have a larger standard deviation—trunk length of a randomly chosen African elephant, or trunk length of a randomly chosen *adult* African elephant?

11. The percentage X of a company's mail orders that require special shipping on any given day is found to have a uniform probability density function, $f(x) = \frac{1}{10}$ for $5 < x < 15$. [f is zero otherwise.]

   (a) Sketch a graph of $f(x)$.

   (b) Find the probability that between 10 and 12 percent of mail orders require special shipping.

   (c) Verify that $\int_{-\infty}^{+\infty} f(x)dx = 1$.

   (d) Find the formula for the cumulative distribution function $F(x)$.

   (e) Find the mean percentage of orders requiring special shipping. Based on the shape of the density curve, can you tell what the median percentage would be?

   (f) Find $Var(X)$.

   (g) Find σ.

12. The time in minutes it takes a PAT bus to get a certain commuter from his bus stop to Pitt is a uniform random variable X with probability density function $f(x) = \frac{1}{4}$ for $12 \leq x \leq 16$.

   (a) Sketch a graph of $f(x)$.

   (b) Find the probability that the commute takes longer than 15 minutes.

   (c) Verify that $\int_{-\infty}^{+\infty} f(x)dx = 1$.

   (d) Find the formula for the cumulative distribution function $F(x)$.

   (e) Find the mean commute time. Based on the shape of the density curve, can you tell what the median commute time would be?

   (f) Find $Var(X)$.

   (g) Find σ.

Solutions to Odd-Numbered Supplemental Exercises A

1. (a) i. $P(X < 2) = f(0) + f(1) = .3 + .4 = .7$

   ii. $P(X \leq 2) = f(0) + f(1) + f(2) = 1$

   iii. $P(X > 1) = f(2) = .3$

   iv. $P(0 < X \leq 2) = f(1) + f(2) = .4 + .3 = .7$

   (b)

x	$F(x)$
0	.3
1	.7
2	1.0

   (c) i. $P(X < 2) = P(X \leq 1) = F(1) = .7$

   ii. $P(X \leq 2) = F(2) = 1$

   iii. $P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - .7 = .3$

   iv. $P(0 < X \leq 2) = P(X \leq 2) - P(X \leq 0) = F(2) - F(0) = 1 - .3 = .7$

3. (a) f would be left-skewed—relatively few people die at a young age; most people die when they are older.

   (b) Because of the left-skewness, the mean would be less than the median.

5. (a) f would be right-skewed—there are unfortunately more murders, suicides, and untimely deaths among younger people.

   (b) Because of the right-skewness, the mean would be greater than the median.

7. (a) f should be normal, as is the case in general for physical characteristics. Therefore it should be symmetric.
   (b) Because of the symmetry of f, the mean should approximately equal the median.
   (c) For a continuous distribution such as this, the probability of any one particular value is zero.

9. Age of Americans has more spread than age of college students, so its standard deviation would be larger.

11. (a) $f(x)$ is a horizontal line of height $\frac{1}{10}$ between $x = 5$ and $x = 15$; it is zero otherwise.
   (b) $P(10 \leq X \leq 12) = \int_{10}^{12} \frac{1}{10} dx = \frac{1}{10} x\Big|_{10}^{12} = \frac{12}{10} - \frac{10}{10} = \frac{2}{10} = .2.$
   (c) $\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{5} 0dx + \int_{5}^{15} \frac{1}{10} dx + \int_{15}^{+\infty} 0dx = 0 + \frac{1}{10}x\Big|_{5}^{15} + 0 = \frac{1}{10}(15 - 5) = 1$
   (d) $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{5}^{x} \frac{1}{10} dt = \frac{1}{10} t\Big|_{5}^{x} = \frac{1}{10}x - \frac{5}{10} = \frac{1}{10}x - \frac{1}{2}$ for x between 5 and 15. $F(x) = 0$ for $x < 5$ and $F(x) = 1$ for $x \geq 15$.
   (e) $\mu = \int_{-\infty}^{+\infty} x f(x)dx = \int_{5}^{15} \frac{1}{10} x dx = \frac{1}{20} x^2\Big|_{5}^{15} = \frac{225}{20} - \frac{25}{20} = \frac{200}{20} = 10.$
   Since f is symmetric, the median must also be 10.
   (f)

$$Var(X) = \left(\int_{-\infty}^{+\infty} x^2 f(x)dx\right) - \mu^2$$

$$= \left(\int_{5}^{15} \frac{1}{10} x^2 dx\right) - 10^2 = \frac{1}{30} x^3 \Big|_{5}^{15} - 100 = \frac{3375}{30} - \frac{125}{30} - 100 = \frac{25}{3}$$

   (g) $\sigma = \sqrt{\frac{25}{3}} = \frac{5}{\sqrt{3}} = 2.89$

Chapter 7 Supplement B

More on Continuous Random Variables and Their Probability Distributions

Recall: To solve for probabilities involving a continuous random variable X, we may either integrate its probability density curve $f(x)$ over the appropriate interval:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

or work with the cumulative distribution function $F(x)$, which is the integral of f:

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad \text{so} \qquad P(a \leq X \leq b) = F(b) - F(a)$$

It follows that f is the derivative of F:

$$f(x) = \frac{d}{dx} F(x)$$

To find the mean, variance, and standard deviation of X, we also need to perform integration. Last time we presented **Formulas for Continuous Random Variables**, which we will utilize when solving for probabilities, means, and standard deviations.

Before working through more examples, we should review some basics for derivatives and integrals of some common elementary functions:

$$\frac{d}{dx} kx^n = nkx^{n-1}$$

$$\int kx^n dx = \frac{k}{n+1} x^{n+1} + C$$

The latter holds for $n \neq -1$. You may recall that $\int x^{-1} dx = \ln x$; however, this result will not be required in our examples.

Example

If $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$, find $f(x)$.

$f(x) = \frac{d}{dx}[3x^2 - 2x^3] = 6x - 6x^2$, also defined for $0 \leq x \leq 1$.

Example

If $f(x) = 6x - 6x^2$ for $0 \leq x \leq 1$, find $P(0 \leq X \leq \frac{1}{2})$.

$$P(0 \leq X \leq \frac{1}{2}) = \int_0^{\frac{1}{2}} 6x - 6x^2 dx$$

$$= (3x^2 - 2x^3)\Big|_0^{\frac{1}{2}} = 3\left(\frac{1}{2}\right)^2 - 2\left(\frac{1}{2}\right)^3 - [3(0)^2 - 2(0)^3] = \frac{3}{4} - \frac{2}{8} = \frac{1}{2}$$

Example

If $f(x) = \frac{1}{8}(x+1)$ for $2 \leq x \leq 4$, find $P(3 \leq X \leq 4)$.

$$P(3 \leq X \leq 4) = \int_3^4 \frac{1}{8}(x+1)dx = \left(\frac{1}{16}x^2 + \frac{1}{8}x\right)\Big|_3^4 = \frac{16}{16} + \frac{4}{8} - \frac{9}{16} - \frac{3}{8} = \frac{7}{16} + \frac{1}{8} = \frac{9}{16}$$

Exponential random variables arise in many situations. Recall that $\frac{d}{dx}e^x = e^x$ and $\int e^x dx = e^x + C$. Because of the chain rule, it follows that

$$\frac{d}{dx}e^{-kx} = -ke^{-kx}$$

$$\int -ke^{-kx}dx = e^{-kx} + C$$

Example

The waiting time X (in days) until a certain brand of watch must be reset is a continuous random variable with probability density function $f(x) = \frac{1}{120}e^{-\frac{x}{120}}$ for $x > 0$. [Otherwise $f(x)$ is zero.] What is the probability of needing to be reset between 1 month (30 days) and 2 months (60 days)?

$$P(30 \leq X \leq 60) = \int_{30}^{60} \frac{1}{120}e^{-\frac{x}{120}} = -e^{-\frac{x}{120}}\Big|_{30}^{60} = -e^{-\frac{60}{120}} + e^{-\frac{30}{120}}$$

$$= e^{-\frac{30}{120}} - e^{-\frac{60}{120}} = .7788 - .6065 = .1723$$

In general, for positive a and b,

$$\int_a^b ke^{-kx}dx = e^{-ka} - e^{-kb}$$

If a is 0, remember that $e^0 = 1$. If b is ∞, remember that $e^{-\infty} = 0$. Using integration by parts, it can be shown that an exponential random variable X with probability density function ke^{-kx} for $x > 0$ and $k > 0$ has the following mean, variance, and standard deviation:

$$E(X) = \mu = \frac{1}{k}; \qquad Var(X) = \sigma^2 = \frac{1}{k^2}; \qquad \sigma = \frac{1}{k}$$

Example

1. For our watch problem, what is the probability of operating up to 84 days until needing to be reset?
   Since $f(x) = \frac{1}{120}e^{-\frac{x}{120}}$ is of the form ke^{-kx} with $k = \frac{1}{120}$, we have
   $$P(0 \leq X \leq 84) = \int_0^{84} \frac{1}{120}e^{-\frac{x}{120}}\,dx = e^{-\frac{1}{120}(0)} - e^{-\frac{1}{120}(84)} \approx 1 - .5 = .5$$
   Thus, $X = 84$ days is the equal-area point of the distribution, or *median.*
2. What are the mean and standard deviation for waiting time until the watch must be reset?
   Since $f(x) = \frac{1}{120}e^{-\frac{x}{120}}$ is of the form ke^{-kx} with $k = \frac{1}{120}$, it follows that $E(X) = \mu = \frac{1}{1/120} = 120$, and also $\sigma = 120$. The waiting time is about 120 days, give or take about 120 days.

Example

Suppose X is a random variable with probability density function $f(x) = \frac{1}{10}e^{-\frac{x}{10}}$ for $x > 0$. Find the mean μ and median M of X.

Since $f(x) = \frac{1}{10}e^{-\frac{x}{10}}$ is of the form ke^{-kx} with $k = \frac{1}{10}$, it follows that $E(X) = \mu = \frac{1}{1/10} = 10$. To find the median M, we set

$$.5 = \int_0^M \frac{1}{10}e^{-\frac{x}{10}}\,dx = -e^{-\frac{x}{10}}\Big|_0^M = -e^{-\frac{M}{10}} + 1$$

So $e^{-\frac{M}{10}} = .5$. Since $e^{-.7} \approx .5$, it follows that $M = 7$.

Note: exponential distributions such as this are right-skewed, with a high probability of being close to zero, tapering down to very low probabilities of extremely large values of x. This results in a mean (10) which is considerably higher than the median (7). Our next example also has a probability density function f that is right-skewed. The problem is stated, however, in terms of the cumulative distribution F.

Example

The shelf-life in hours of a certain perishable packaged food is a random variable X with cumulative distribution function

$$F(x) = 1 - 10,000(x + 100)^{-2} \quad \text{for} \quad x \geq 0;$$

[$F(x)$ is zero for $x < 0$. Note that shelf-lives can't be negative.]

1. Find the probability density function $f(x)$.
2. Use f to find the probability that a randomly chosen package will last
   (a) less than 41 hours
   (b) less than 100 hours
   (c) more than 200 hours
   (d) between 100 and 200 hours
3. Verify that $\int_{-\infty}^{+\infty} f(x)dx = 1$.
4. Re-solve the probabilities in part (2), using F.
5. Find $E(X) = \mu$.

Here are the solutions:

1. $f(x) = \frac{d}{dx}F(x) = 20,000(x + 100)^{-3}$ for $x \geq 0$, 0 otherwise.

2. Here we use f to solve for probabilities:

(a)

$$P(X < 41) = \int_0^{41} 20,000(x+100)^{-3}dx = -10,000(x+100)^{-2}\Bigg|_0^{41}$$

$$= -10,000(141)^{-2} + 10,000(100)^{-2} = -\frac{10,000}{19,881} + \frac{10,000}{10,000} \approx -\frac{1}{2} + 1 = \frac{1}{2} = .5$$

Thus, 41 days is the *median* shelf-life; the area under $f(x)$ is .5 on either side of $x = 41$.

(b)

$$P(X < 100) = \int_0^{100} 20,000(x+100)^{-3}dx = -10,000(x+100)^{-2}\Bigg|_0^{100}$$

$$= -10,000(200)^{-2} + 10,000(100)^{-2} = -\frac{10,000}{40,000} + \frac{10,000}{10,000} = -\frac{1}{4} + 1 = \frac{3}{4}.$$

Thus, $\frac{3}{4}$ of the area under f is to the left of $x = 100$.

(c)

$$P(X > 200) = \int_{200}^{\infty} 20,000(x+100)^{-3}dx = -10,000(x+100)^{-2}\Bigg|_{200}^{\infty}$$

$$= \lim_{x\to\infty} -10,000(x+100)^{-2} + 10,000(300)^{-2} = 0 + \frac{10,000}{90,000} = \frac{1}{9}$$

[Note that as x gets infinitely large, $(x+100)^{-2} = \frac{1}{(x+100)^2}$ goes to zero.]

(d)

$$P(100 < X < 200) = \int_{100}^{200} 20,000(x+100)^{-3} = -10,000(x+100)^{-2}\Bigg|_{100}^{200}$$

$$= -10,000(300)^{-2} + 10,000(200)^{-2} = -\frac{1}{9} + \frac{1}{4} = \frac{-4+9}{36} = \frac{5}{36}$$

3.

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^{\infty} 20,000(x+100)^{-3}dx = -10,000(x+100)^{-2}\Bigg|_0^{\infty} = 0 + \frac{10,000}{10,000} = 1$$

4. Now we use F to solve for probabilities:

(a)

$$P(X < 41) = P(X \leq 41) = F(41) = 1 - 10,000(41+100)^{-2} = 1 - \frac{10,000}{19,881} \approx 1 - \frac{1}{2} = \frac{1}{2}$$

(b)

$$P(X < 100) = P(X \leq 100) = F(100) = 1 - 10,000(100+100)^{-2} = 1 - \frac{10,000}{40,000} = 1 - \frac{1}{4} = \frac{3}{4}$$

(c)

$$P(X > 200) = 1 - P(X \leq 200) = 1 - F(200) = 1 - [1 - 10,000(200+100)^{-2}] = \frac{10,000}{90,000} = \frac{1}{9}$$

(d)

$$P(100 < X < 200) = F(200) - F(100) = [1 - 10,000(200+100)^{-2}] - [1 - 10,000(100+100)^{-2}]$$

$$= \frac{8}{9} - \frac{3}{4} = \frac{32-27}{36} = \frac{5}{36}$$

5.

$$E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^{\infty} 20,000x(x+100)^{-3}dx$$

This type of integral may be solved using the technique of substitution, or with integration by parts. These skills will not be required in this course, but solutions are presented here for students who wish to review.

(a) Using substitution to simplify the expression being raised to a power:
Let $u = x + 100$. Then $x = u - 100$, and $\frac{du}{dx} = 1$ means $dx = du$. We adjust the limits of integration as follows: when $x = 0$, $u = 100$; as $x \to \infty$, $u \to \infty$.

$$\int_0^{\infty} 20,000x(x+100)^{-3}dx = \int_{100}^{\infty} 20,000(u-100)u^{-3}du = \int_{100}^{\infty} (20,000u^{-2} - 2,000,000u^{-3})du$$

$$= -20,000u^{-1}\Big|_{100}^{\infty} + 1,000,000u^{-2}\Big|_{100}^{\infty}$$

$$= \lim_{u\to\infty} \frac{-20,000}{u} + 20,000(100)^{-1} + \lim_{u\to\infty} \frac{1,000,000}{u^2} - 1,000,000(100)^{-2}$$

$$= 0 + 200 + 0 - 100 = 100$$

The mean shelf-life is 100 hours. Because of the extreme right-skewness of the density function $f(x)$, the mean is much larger than the median, 41.

(b) Using the integration by parts formula $\int udv = uv - \int vdu$:
Take $u = 20,000x$, $dv = (x+100)^{-3}dx$, $du = 20,000dx$, $v = -\frac{1}{2}(x+100)^{-2}$. Then

$$E(X) = -10,000x(x+100)^{-2}\Big|_0^{\infty} + \int_0^{\infty} 10,000(x+100)^{-2}dx$$

$$= -10,000 \lim_{x\to\infty} \frac{x}{(x+100)^2} + 10,000(0)(0+100)^{-2} - 10,000(x+100)^{-1}\Big|_0^{\infty}$$

$$= 0 + 0 - 10,000 \lim_{x\to\infty} \frac{1}{x+100} + 10,000(0+100)^{-1} = \frac{10,000}{100} = 100$$

Similar calculations can be used to find the variance and standard deviation. They turn out to be infinite, which may not be entirely realistic, but we must remember that such models are merely idealizations which provide fairly reasonable answers to most questions.

Important: Writing “french” instead of “French” on a history paper is just a minor mistake. Confusing “$f(x)$” (the probability density function) and “$F(x)$” (the cumulative distribution function) in statistics is a major mistake. In a sense, they are opposites: F is the integral of f, f is the derivative of F.

Supplemental Exercises B

1. A random variable X has cumulative distribution function $F(x) = \frac{1}{16}(x^2 + 2x - 8)$ for $2 < x < 4$. Find the probability that X takes a value

   (a) less than 2.5

(b) greater than 3

(c) between 2.5 and 3

(d) less than 1

(e) less than 5

2. A random variable X has cumulative distribution function $F(x) = 3x^2 - 2x^3$ for $0 < x < 1$. Find the probability that X takes a value

(a) less than .5

(b) greater than .6

(c) between .5 and .6

(d) less than -.5

(e) less than 2

3. Find the probability density function $f(x)$, given the cumulative distribution function. Be sure to specify the interval where f is defined.

(a) $F(x) = 21x^{20} - 20x^{21}$ for $0 < x < 1$

(b) $F(x) = 1 - e^{-\frac{1}{2}x}$ for $0 \leq x < \infty$

(c) $F(x) = 1 - 9x^{-2}$ for $x > 3$

4. Find the probability density function $f(x)$, given the cumulative distribution function. Be sure to specify the interval where f is defined.

(a) $F(x) = \frac{2}{3}x^2 - \frac{8}{27}x^3 + \frac{1}{27}x^4$ for $0 \leq x \leq 3$

(b) $F(x) = 1 - e^{-3x}$ for $x \geq 0$

(c) $F(x) = 1 - x^{-3}$ for $x \geq 1$

5. Find the cumulative distribution function $F(x)$, given the probability density function $f(x) = \frac{1}{8}$ for $-4 < x < 4$. Be sure to specify each interval where F is defined.

6. Find the cumulative distribution function $F(x)$, given the probability density function $f(x) = \frac{1}{2}$ for $-1 < x < 1$. Be sure to specify each interval where F is defined.

7. The weight X, in pounds, of a manufactured product is a random variable with probability density function $f(x) = \frac{4}{15}x^3$ for $1 < x < 2$.

(a) Find the probability of such a product weighing less than 1.71 pounds. Explain why 1.71 is (approximately) the median weight.

(b) Verify that $\int_{-\infty}^{+\infty} f(x)dx = 1$.

(c) Find the mean weight μ.

(d) Comparing your answers to (a) and (c), do you think a graph of $f(x)$ would exhibit left-skewness or right-skewness?

(e) Find $Var(X)$ and σ. As far as the spread is concerned, do you think this is a realistic model for the weight of cereal boxes that are supposed to be about 1.7 pounds?

8. A certain charity is planning a direct-mail campaign. The fraction X of non-respondents is assumed to have probability density function $f(x) = 20x^3 - 20x^4$ for $0 \leq x \leq 1$.

(a) Find the probability that the proportion of non-respondents is less than .69. Explain why .69 is (approximately) the median.

(b) Verify that $\int_{-\infty}^{+\infty} f(x)dx = 1$.

(c) Find the expected proportion of non-respondents. Does the charity have good reason to hope that a majority of those who received their mailing *will* respond?

(d) Which is larger, the mean or the median?

(e) Find $Var(X)$ and σ.

9. The tread wear X (in thousands of kilometers) for a certain brand of tire has probability density function $f(x) = \frac{1}{30}e^{-\frac{x}{30}}$ for $x > 0$.

(a) What is the probability of lasting between 27 and 36 thousand kilometers?

(b) What is the expected tread wear of such a tire?

(c) What is the standard deviation of tread wear?

10. The reservations manager for a large airline finds that the waiting time X in minutes between phone calls to the reservation center has probability density function $f(x) = 2e^{-2x}$ for $x \geq 0$.

(a) What is the probability that the waiting time will be between 1 and 2 minutes?

(b) What is the mean waiting time?

(c) What is the standard deviation of waiting time?

(d) What is the median waiting time?

Solutions to Odd-Numbered Supplemental Exercises B

1. (a) $P(X < 2.5) = F(2.5) = \frac{1}{16}(2.5^2 + 2(2.5) - 8) \approx .20$

(b) $P(X > 3) = 1 - P(X \leq 3) = 1 - F(3) = 1 - \frac{1}{16}(3^2 + 2(3) - 8) \approx 1 - .44 = .56$

(c) $P(2.5 < X < 3) = F(3) - F(2.5) \approx .44 - .20 = .24$

(d) $P(X < 1) = 0$ since we can assume F to be 0 below $x = 2$.

(e) $P(X < 5) = 1$ since X must be between 2 and 4.

3. (a) $f(x) = 420x^{19} - 420x^{20}$ for $0 < x < 1$

(b) $f(x) = \frac{1}{2}e^{-\frac{1}{2}x}$ for $0 \leq x < \infty$

(c) $f(x) = 18x^{-3}$ for $x > 3$.

5. $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-4}^{x} \frac{1}{8}dt = \frac{1}{8}t\Big|_{-4}^{x} = \frac{1}{8}x + \frac{1}{2}$ for $-4 < x < 4$. $F(x)$ is zero for $x \leq 4$, one for $x \geq 4$.

7. (a) $P(X < 1.71) = \int_{1}^{1.71} \frac{4}{15}x^3 dx = \frac{1}{15}x^4\Big|_{1}^{1.71} \approx .57 - .07 = .50$. Since X is less than 1.71 approximately half of the time, 1.71 is roughly the median.

(b) $\int_{-\infty}^{+\infty} f(x)dx = \int_{1}^{2} \frac{4}{15}x^3 dx = \frac{1}{15}x^4\Big|_{1}^{2} = \frac{16-1}{15} = 1$

(c) $\mu = \int_{-\infty}^{+\infty} xf(x)dx = \int_{1}^{2} \frac{4}{15}x^4 dx = \frac{4}{75}x^5\Big|_{1}^{2} = \frac{128-4}{75} \approx 1.653$

(d) The mean 1.653 is less than the median 1.71; the distribution is somewhat left-skewed.

(e)

$$Var(X) = \left(\int_{-\infty}^{+\infty} x^2 f(x)dx\right) - \mu^2$$

$$\approx \int_{1}^{2} \frac{4}{15}x^5 dx - (1.653)^2 \approx \frac{2}{45}x^6 \Bigg|_{1}^{2} - 2.73 = 2.8 - 2.73 = .07 \quad \text{pounds}$$

$\sigma = \sqrt{.07} = .26$ pounds, about one fourth of a pound, or 4 ounces. This seems high for a product like boxes of cereal, whose weight should conform more closely to the target weight.

9. Note that $f(x)$ is of the form ke^{-kx} with $k = \frac{1}{30}$.

(a)

$$P(27 < X < 36) = \int_{27}^{36} \frac{1}{30} e^{-\frac{x}{30}} dx = e^{-\frac{1}{30}(27)} - e^{-\frac{1}{30}(36)} \approx .41 - .30 = .11$$

(b) $\mu = \frac{1}{k} = \frac{1}{1/30} = 30$ thousand kilometers

(c) $\sigma = \frac{1}{k} = 30$ thousand kilometers.