

## Chapter 3 – Mathematically Representing Probability Distributions Using Random Variables

### 3.1 - Introduction

In previous chapters we developed techniques for describing events and for quantifying the probability of their occurrence. These principles were applied to a variety of specific engineering problems. In this chapter we will unify our approach by developing mathematical tools to quantify probabilities of occurrence of events. This will provide an analytical basis for Chapter 4, where we will show that the myriad of problems that we might encounter in engineering design can be grouped into a few broad categories of problem types that can be represented by algebraically distinct probability distributions.

### 3.2 - Random Variables

A *random variable* is a mathematical quantity that can be assigned to represent a probabilistic event. Many of the events whose probabilities we wish to determine fall quite naturally into a quantitative framework for description by a random variable. An example would be the dump trucks of Example Problem 2.2. We could define events as follows:

- $E_0 \rightarrow$  no trucks operate on a given day
- $E_1 \rightarrow$  1 truck operates on a given day
- $E_2 \rightarrow$  2 trucks operate on a given day
- $E_3 \rightarrow$  3 trucks operate on a given day

If we define a random variable,  $X$ , to be the number of trucks operating on a given day, then the values of the random variable,  $X = 0, 1, 2,$  or  $3$ , have a one-to-one correspondence with the events  $E_0, E_1, E_2,$  and  $E_3$ , respectively. The advantage of the random variable approach to designating the events is that whereas  $E_0, E_1$  etc. fit nicely into the English language, the idea of  $X = 0, 1,$  etc. fits better into the language of mathematics. For example, it is very easy to fit the events described by random variables into the uniaxial sample spaces described in Section 1.6. Figure 3.2.1 illustrates this for the truck example.

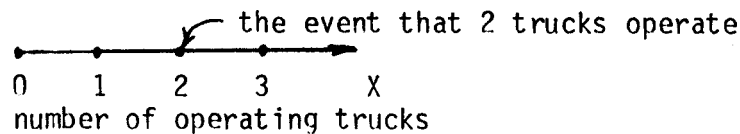


Figure 3.2.1 Sample space depicting the events 0, 1, 2, or 3 trucks operating in terms of random variable,  $X$ .

The procedure that we use to manipulate physical processes so that they can be described by a random variable is essentially identical to the process used in Section 1.6 to manipulate physical processes into a uniaxial sample space. Some processes do not fall into place as naturally as did the dump trucks example, but they can be set in the context of a random variable with a little imagination. For instance, to describe the position of a person in line, the first person in line could be associated with  $X = 1$ , the second person with  $X = 2$ , etc. This same principle applies to events that fit onto tabular sample spaces. If the extended Venn diagram describing a sample space consists of a one column table, the first entry into the table can be assigned to  $X = 1$ , the second entry to  $X = 2$ , etc.

Still, other physical processes must be blatantly forced into the context of a random variable. If a warehouse has beam types W14x119, PH12-74, and S24x120, we could let  $X = 1$  denote W14x119's;  $X = 2$  denote HP12x74's; and  $X = 3$  denote S24x120's.

### 3.3 - Probability distributions for discrete events - the probability mass function.

Having associated specific events with a random variable, it is now possible to mathematically represent the way in which the probability of occurrence of an event varies with the value of the random variable characterizing the event. In Example Problem 2.2, the probability of no trucks operating was 0.125. Thus,  $P(E_0) = 0.125$ . In the random variable framework, this will be written as

$$P(X = 0) = 0.125$$

We can generalize this notation as follows:

$$P(X = x) = k \tag{3.3.1}$$

which states that the probability that the random variable,  $X$ , takes on the specific value,  $x$ , is numerically equal to  $k$ .

The probability of occurrence for each and every possible value of  $X$  is given by the *probability mass function* (PMF). For Example Problem 2.2 the PMF is:

$$P(X = 0) = 0.125 \tag{3.3.2a}$$

$$P(X = 1) = 0.375 \tag{3.3.2b}$$

$$P(X = 2) = 0.375 \tag{3.3.2c}$$

$$P(X = 3) = 0.125 \tag{3.3.2d}$$

Note that this PMF is composed of four distinct equations taken as a group, and the sum of the probabilities is 1.0 because the events are mutually exclusive and collectively exhaustive.

It is possible to graphically represent the PMF by adding a second axis (ordinate) to the uniaxial extended Venn diagram. The probability of occurrence is plotted on this ordinate for each value of the discrete random variable. Figure 3.3.1 shows how Figure 3.2.1 can be extended to graphically represent Eq. (3.3.2). Graphs like Figure 3.3.1 are also referred to as PMF's for discrete random variables.

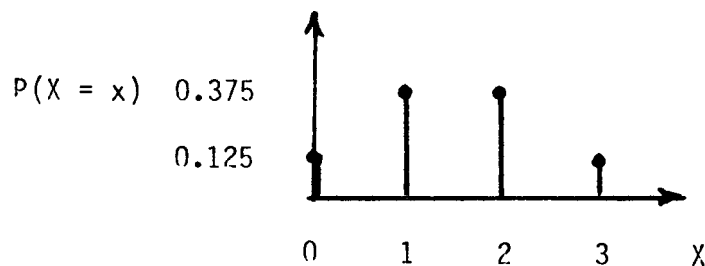


Figure 3.3.1 Graphically depicting the probability mass function (PMF) given by Eq. (3.3.2).

Equation (3.3.2) and Fig. 3.3.1 give a comprehensive view of how the probability of occurrence of events varies with all possible situations relative to the number of trucks operating on a given day. Eq. (3.3.2) presents this in a mathematically tractable form. Figure 3.3.1 presents the same in a visually convenient form. Thus, the random variable approach is a useful approach if you want to produce a comprehensive representation of the big picture about a statistical process, rather than just trying to calculate the probability of some isolated event.

In order to construct a PMF for a problem, two criteria must be met. First, the sample space must be discrete. Second, an event corresponding to a particular value of the random variable must be mutually exclusive from events corresponding to every other value of the random variable. This is because each value of the random variable, being discrete, cannot overlap (intersect with) any other value of the random variable.

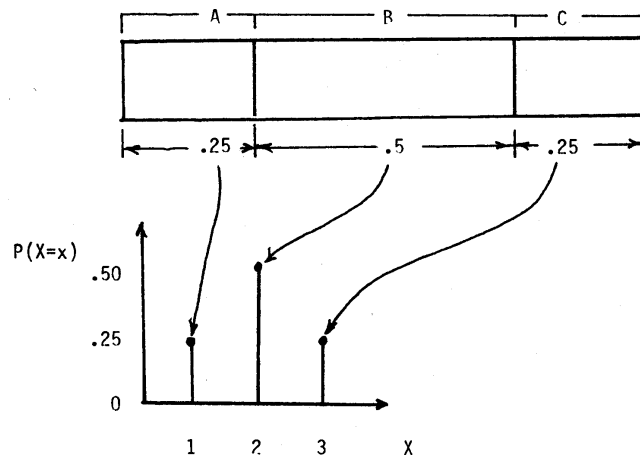
Sometimes, converting a sample space to a PMF is straightforward. Example Problem 3.1 shows how the same sample space shown in Fig. 2.2.1 can be easily represented by a random variable. Sometimes we must exercise care to insure that mutually exclusive events are constructed to facilitate creating a PMF. Example Problem 3.2 shows how this can be done for the more complicated sample space of Fig. 2.2.2.

### EXAMPLE PROBLEM 3.1

#### Constructing a PMF for a sample space of mutually exclusive events

Consider the sample space shown in Fig. 2.2.1 to which we have added event C such that A, B, and C are mutually exclusive and collectively exhaustive. Recall that  $P(A) = .25$ ,  $P(B) = .50$ , and the student who completed the exercise would have found that  $P(C) = .25$ .

We can create a random variable X such that  $X = 1$  corresponds to A;  $X = 2$  corresponds to B; and  $X = 3$  corresponds to C. Thus,  $P(1) = .25$ ,  $P(2) = .50$ , and  $P(3) = .25$ . The figure below demonstrates how the Venn diagram can be converted into a PMF.



Show  $A \cup B$  on the PMF and calculate  $P(A \cup B)$ .

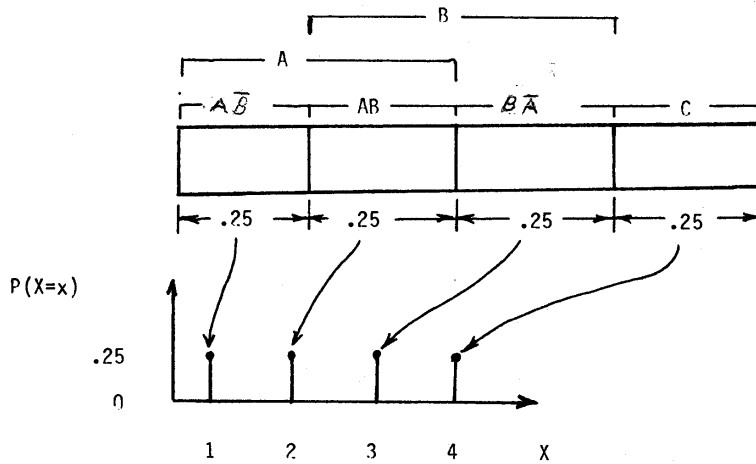
Show S on the PMF and calculate  $P(S)$ .

### EXAMPLE PROBLEM 3.2

#### Constructing a PMF for a sample space containing non-mutually exclusive events

Consider the sample space shown in Fig. 2.2.2. Recall that  $P(A) = .50$ ,  $P(B) = .50$ , and  $P(AB) = .25$ . For completeness, we will add an event  $C$  such that events  $A$ ,  $B$ , and  $C$  are collectively exhaustive.  $P(C)$  is equal to  $.25$ .

The difficulty in transforming this sample space into a PMF lies in the fact that  $A$  and  $B$  intersect (are not mutually exclusive). To overcome this, let  $X = 1$  correspond to  $A\bar{B}$ ;  $X = 2$  correspond to  $AB$ ;  $X = 3$  correspond to  $\bar{B}A$ ; and  $X = 4$  correspond to  $C$ . We then have 4 mutually exclusive, collectively exhaustive events that can be mapped into a discrete random variable. The figure below shows the PMF for this problem.



Identify events  $A$ ,  $B$ , and  $A \cup B$  on the PMF. Calculate their probabilities.  
 Identify  $S$  on the PMF and calculate its probability.

### 3.4 - The Cumulative Distribution Function for Discrete Random Variables

Sometimes it is useful to be able to quantify the probability that a random variable takes on values less than or equal to some particular value. Thus, we may be interested in the probability that there will be two or fewer trucks operating on a given day. This event would be  $E = E_0 \cup E_1 \cup E_2$ , and because the events are mutually exclusive,  $P(E) = P(E_0) + P(E_1) + P(E_2)$ . In random variable notation this same calculation would be written as

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) \quad (3.4.1)$$

We can generalize this by defining the *cumulative distribution function* (CDF),  $F_X(x)$ , to be

$$F_X(x) \equiv P(X \leq x) = \sum_{\text{all } x_i \leq x} P(X = x_i) \quad (3.4.2)$$

Figure 3.4.1 shows the CDF for the trucks problem whose PMF was shown in Figure 3.3.1.

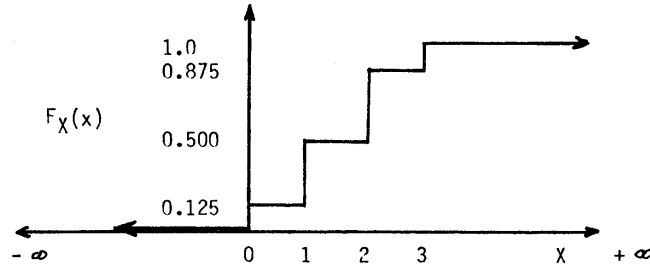


Figure 3.4.1 Cumulative distribution function (CDF) for Example Problem 2.2

Note that the CDF is not limited to integer values of  $X$  as was the PMF. Rather, it is a continuous function of  $X$ . Moreover, it is not confined to the range of values of  $X$  within the sample space. Rather, it is defined for  $-\infty \leq X \leq +\infty$ . It has the following additional properties:

- 1)  $F_X(-\infty) = 0$ ,
- 2)  $F_X(+\infty) = 1.0$ ,
- 3)  $F_X(x) \geq 0$ ,
- 4)  $F_X(x)$  always increases with increasing  $X$ , and
- 5)  $F_X(x)$  is continuous.

The student should construct CDF's for the PMF's developed in Example Problems 3.1 and 3.2.

The CDF is useful partly because it is a means for communicating the probability that  $X$  is less than or equal to a certain value. However, it is also useful because we can apply the concept of the CDF to continuous random variables in a manner that circumvents some mathematical difficulties. This is accomplished in the next section.

### 3.5 - Probability distributions for continuous random variables: the cumulative distribution function and the density function

A conceptual difficulty arises when we try to extend the concepts relating to PMF's to continuous random variables. Consider the strength of a concrete test cylinder. At first glance it seems easy enough to calculate the probability that a cylinder would exhibit a strength of, say, 2000 psf. In actuality,  $P(\text{strength} = 2000 \text{ psf}) = 0$ . The reason for this is that any particular cylinder would have a strength of 2000.0001, or perhaps 1999.9997; but we would never find one with a strength of exactly 2000.0000000---.

For a continuous random variable, we can only assign a probability if we are given a range of values for the random variable. Thus it would be possible to calculate the probability that the strength of a cylinder is greater than 1990 pfs but less than or equal to 2010 psf. The cylinders that tested at 2000.0001 and at 1999.9997 would both fall within this range.

Given that we must deal with defining the probability over a range of values for a continuous random variable, it is mathematically convenient to look at an interval extending from some

particular value of  $X$  back to  $X = -\infty$ . This approach is consistent with the definition previously presented for a CDF. Thus, for a continuous random variable we still define the CDF as

$$F_X(x) = P(X \leq x) \quad (3.5.1)$$

Conceptually, we can graph this function in a manner similar to Fig. 3.4.1; however, because our random variable is continuous rather than discrete, the CDF will be smooth rather than step-wise. Such a function is shown in Fig. 3.5.1:

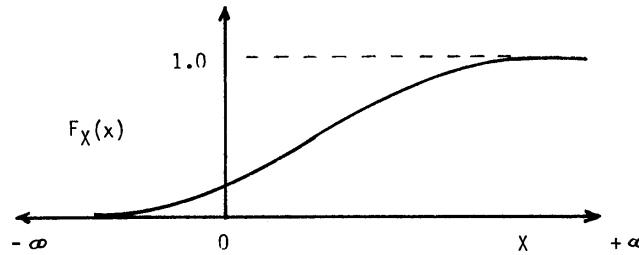


Figure 3.5.1 CDF for a hypothetical continuous random variable.

This curve obeys all of the rules set out in Section 3.4 for the CDF for a discrete random variable; therefore, it is logical to use the notation  $F_X(x)$  for CDF's for both discrete and continuous random variables.

Now, if the CDF at  $X = x$  for a discrete random variable was obtained by summing over the PMF for the discrete random variable for  $X \leq x$  (Eq. 3.4.2), then for a continuous random variable there must exist some function  $f_X(x)$  such that we can obtain  $F_X(x)$  by integrating  $f_X(x)$  from  $-\infty$  to  $x$ . Thus,

$$F_X(x) = \int_{-\infty}^x f_X(\zeta) d\zeta \quad (3.5.2)$$

Here,  $\zeta$  is a dummy variable of integration and vanishes when the definite integral is evaluated; thus,  $F_X(x)$  is indeed a function of  $x$  only. The function  $f_X(x)$  is termed the *probability density function* (PDF). It occupies the same position with respect to continuous random variables that the PMF occupies with respect to discrete random variables. The summation operation reflected in Eq. (3.4.2) for discrete random variables is replaced by the integration operation reflected in Eq. (3.5.2) for continuous random variables. Because the CDF for continuous random variables is the integral of the PDF, we can obtain the PDF by differentiating the CDF with respect to the random variable,  $X$

$$f_X(x) = \frac{dF_X(x)}{dx} \quad (3.5.3)$$

Thus, the value of the PDF at  $X = x$  is numerically equal to the slope of the CDF at  $X = x$ .

From a practical point of view, it is possible to generate the CDF for a continuous random variable. We can use the method of empirical observations to determine the probability that  $X \leq x$  for the entire range of the random variable  $X$ . Then we can fit a smooth curve through this relationship to generate the CDF. Finally, we can take the derivative of the CDF (either

analytically or numerically) according to Eq. (3.5.3) to generate the PDF for the random variable. Example Problem 3.3 demonstrates this process for a sample data set, however, in practical situations we will use alternate approaches to be described in Chapter 5.

It is possible to rationalize the concept of PDF for a continuous random variable by setting an analogy with another similar but more common concept—that of density of an engineering material. Consider a block of steel. At each point this steel has a density of 450 pcf. However, the weight of the steel at a point is zero because a point has no volume. One can create a weight only by opening up a space around the point to consider a finite volume.

This same analogy extends to the concept of a probability density function, and explains the reason for the inclusion of the word “density” in the name of the PDF. At a particular point on the X axis, we can define the density of the probability (assign a value to the PDF). However, the probability of X taking on that exact value is zero because we have not opened up any region (volume) within the sample space. We only open up a volume when we examine  $P(a < X \leq b)$ . The PDF reflects the potential for a finite probability to be displayed at a specific value of  $X = x$  provided that we are willing to open up a little volume in the sample space around  $X = x$ .

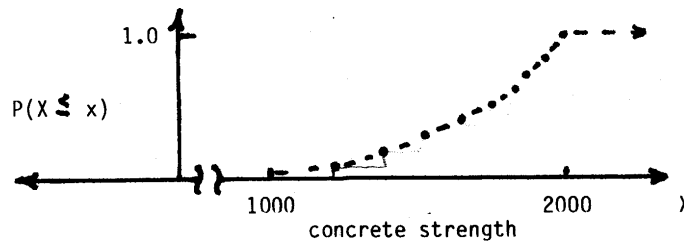
### EXAMPLE PROBLEM 3.3

#### Developing a PDF for a continuous random variable

The following list contains the results of strength tests on concrete cylinders presented in order of increasing strength: 1230, 1390, 1530, 1660, 1750, 1810, 1870, 1920, 1960, and 2000 psf. The following table uses the method of empirical observations to show the probability that the random variable, X, is less than or equal to a certain value, x, over the range of the test values. The probabilities were computed by taking the total number of observations showing strengths less than or equal to x and dividing that number by the total number of tests (10).

strength value, x	1230	1390	1530	1660	1750	1810	1870	1920	1960	2000
test no. i	1	2	3	4	5	6	7	8	9	10
$P(X \leq x)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

The figure below plots the CDF based on the values in the above table.



### EXAMPLE PROBLEM 3.3 (Cont.)

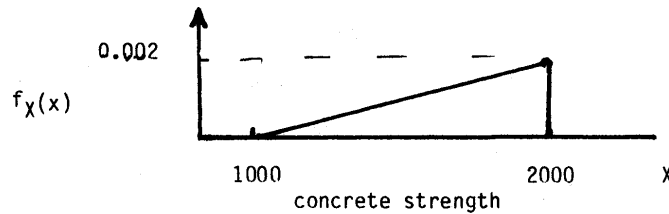
An astute mathematician analyzing the data decided that it was represented by a parabola whose equation is

$$\begin{aligned} F_X(x) &= 1.00 \times 10^{-6} x^2 - 0.002 x + 1.00 = 0 & \text{for } 1000 < X < 2000 \\ F_X(x) &= 0 & \text{for } X \leq 1000 \\ F_X(x) &= 1.0 & \text{for } X \geq 2000 \end{aligned}$$

This relationship is shown as the dotted line on the above plot. We can use Eq. (3.5.3) to obtain the PDF

$$\begin{aligned} f_X(x) &= \frac{dF_X(x)}{dx} = 2.00 \times 10^{-6} x - 0.002 = 0 & \text{for } 1000 < X \leq 2000 \\ f_X(x) &= 0 & \text{elsewhere.} \end{aligned}$$

This relationship is plotted below.



### 3.6 - Calculating the Probability that a Random Variable Lies between Two Values

The CDF provides a convenient means, not only to determine the probability that a random variable is less than or equal to some particular value, but also to determine the probability that the random variable lies between two values. In general, we can write

$$P(a < X \leq b) = \int_a^b f_X(x) dx = F_X(b) - F_X(a) \quad (3.6.1)$$

Thus, to determine the probability that  $X$  lies between  $a$  and  $b$ , we need only to evaluate the CDF at  $b$  and  $a$  and take the difference. This concept also holds for discrete distributions.

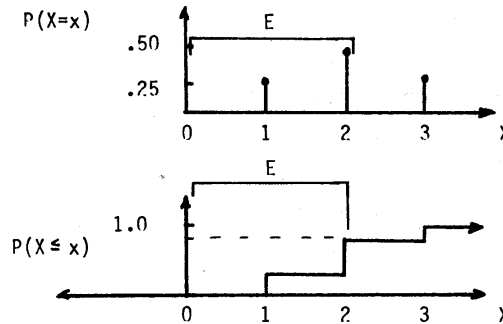
$$P(a < X \leq b) = \sum_{\substack{x_i \leq b \\ x_i > a}} P(X = x_i) = F_X(b) - F_X(a) \quad (3.6.2)$$

Example Problem 3.4 demonstrates the use of Eq. (3.6.1) when applied to Example Problem 3.1 and Example Problem 3.5 demonstrates the use of Eq. (3.6.2) when applied to Example Problem 3.3.

### EXAMPLE PROBLEM 3.4

Calculating the probability that a discrete random variable lies within an interval

Calculate the probability that the random variable of Example Problem 3.1 is greater than 0 but less than or equal to 2. The figure below shows the PMF and CDF for the random variable.



From Eq. (3.6.1)

$$P(0 < X \leq 2) = F_X(2) - F_X(0) = 0.75 - 0.0 = 0.75$$

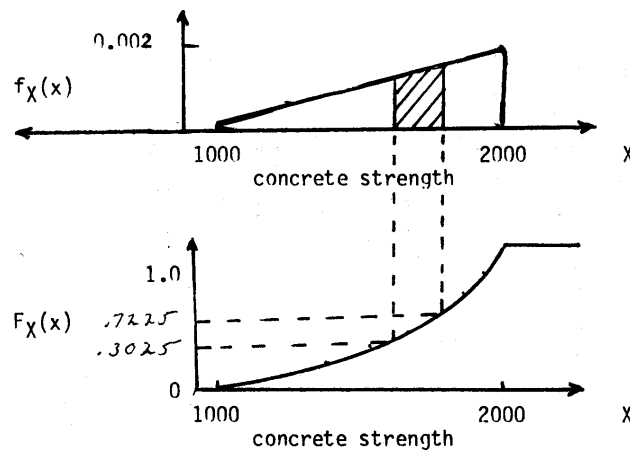
Note that the event referred to consists of  $A \cup B$ . We could also calculate the probability of this event as follows:

$$P(0 < X \leq 2) = P(X = 1) + P(X = 2) = 0.25 + 0.50 = 0.75$$

### EXAMPLE PROBLEM 3.5

Calculating the probability that a continuous random variable lies between two values

Calculate the probability that the random variable of Example Problem 3.3 lies between 1550 and 1850 psf. The figure below shows the PDF and CDF for the problem.



### EXAMPLE PROBLEM 3.5 (Cont.)

The event of interest is shown cross-hatched in the above figure. Applying Eq. (3.6.1) we see that:

$$P(1550 < X \leq 1850) = F_X(1850) - F_X(1550) = 0.7225 - 0.3025 = 0.42$$

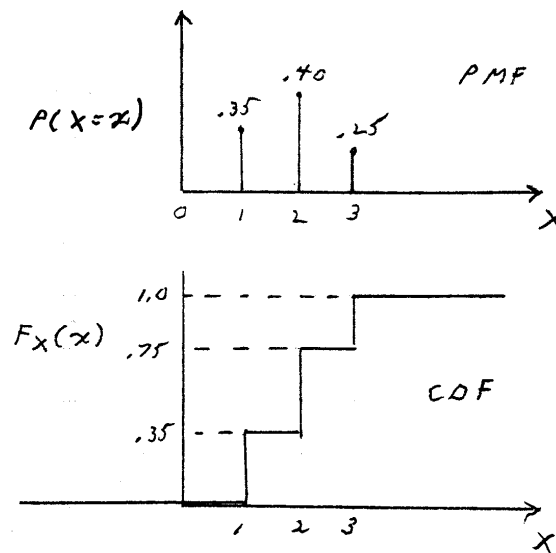
### EXAMPLE PROBLEM 3.6

Using PMFs, PDFs and CDFs

Concrete mix trucks can take any one of three routes to a job site. Taking the tollway requires 40 minutes, taking the expressway requires 60 minutes, and taking Main Street requires 90 minutes. The probability that the driver will select the tollway is 0.35, the probability of his selecting the expressway is 0.40, and the probability of his selecting Main Street is 0.25.

(a) Represent this problem as a random variable. Construct its PMF and CDF.

Solution. Let  $X$  be the random variable representing the route taken, and arbitrarily assign  $X = 1$  to tollway,  $X = 2$  to expressway, and  $X = 3$  to Main Street. Thus,  $P(X = 1) = 0.35$ ,  $P(X = 2) = 0.40$ , and  $P(X = 3) = 0.25$ . The PMF and CDF are shown below:



EXAMPLE PROBLEM 3.6 (Cont.)

(b) Using random variable notation, describe the event of travel time exceeding one hour and calculate its probability.

Solution. The event of travel time exceeding one hour is  $X = 3$ .  $P(X = 3) = 0.25$ .

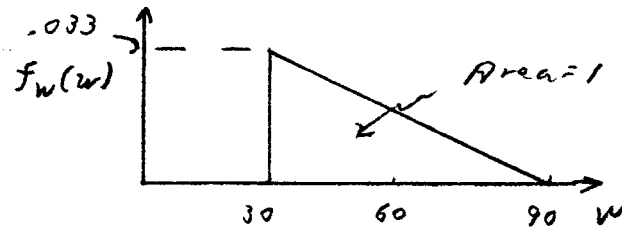
Problem statement continued. Once the mixer reaches the job site, the waiting time for unloading has a triangular distribution dropping from a high at 30 minutes to zero at 90 minutes.

(c) Express this in random variable notation and write equations for and plot the PDF and CDF for waiting time.

Solution. The area under the PDF must be 1.0, therefore the height of the triangle can be found:

$$\frac{1}{2}(90 - 30)h = 1.0 \quad \therefore h = \frac{1}{30}$$

Let the random variable,  $W$ , represent waiting time. Its PDF is shown graphically as:



The equation of the PDF is:

$$f_W(w) = 0 \quad \text{for } W < 30 \text{ and } W > 90$$

$$f_W(w) = \frac{1}{30} \left( -\frac{w}{60} + 1.5 \right) \quad \text{for } 30 \leq W \leq 90$$

The CDF is the integral of the PDF. However, because the PDF is defined in three intervals, we must be careful with the integration process.

For  $W < 30$ , the CDF is 0.

$$\begin{aligned} \text{For } 30 \leq W \leq 90, \quad F_W(w) &= \int_{30}^w \frac{1}{30} \left( -\frac{\zeta}{60} + 1.5 \right) d\zeta = \left( -0.000275\zeta^2 + 0.0495\zeta \right) \Big|_{30}^w \\ &= -0.000275w^2 + 0.0495w - 1.2375 \end{aligned}$$

For  $W > 90$ ,  $F_W(w) = 1.0$

EXAMPLE PROBLEM 3.6 (Cont.)

(d) Express the event waiting time will be between 60 and 80 minutes in random variable notation and calculate its probability.

$$\text{Solution. } P(60 \leq W \leq 80) = \int_{60}^{80} f_W(w) dw = \int_{60}^{80} \frac{1}{30} \left( -\frac{w}{60} + 1.5 \right) dw = 0.22$$

$$\text{Or } P(60 \leq W \leq 80) = F_W(80) - F_W(60) = 0.22$$

(e) Calculate the probability that a mix truck will travel to the site and get unloaded within 100 minutes. Assume waiting time and travel mode are statistically independent.

Solution. Let the random variable, T, represent total time, and use the Theorem of Total Probability:

$$P(T \leq 100) = P(T \leq 100 | X = 1)P(X = 1) + P(T \leq 100 | X = 2)P(X = 2) + P(T \leq 100 | X = 3)P(X = 3)$$

But,

$$P(T \leq 100 | X = 1) = P(W \leq 60) = 0.75$$

$$P(T \leq 100 | X = 2) = P(W \leq 40) = \frac{11}{36}$$

$$P(T \leq 100 | X = 3) = P(W \leq 10) = 0$$

Thus,

$$P(T \leq 100) = 0.75 \cdot P(X = 1) + \frac{11}{36} \cdot P(X = 2) + 0 \cdot P(X = 3) = 0.75 \cdot 0.35 + \frac{11}{36} \cdot .40 + 0 \cdot .25 = 0.385$$

(f) If the mix truck is known to have gotten to the job and to have gotten unloaded within 100 minutes, calculate the probability that it went via the expressway.

$$\text{Solution. } P(X = 2 | T \leq 100) = \frac{P(X = 2 \cap T \leq 100)}{P(T \leq 100)} = \frac{\frac{11}{36}(0.40)}{0.385} = 0.317 \text{ (see Venn diagram below)}$$

