Topic 7

Random Variables and Distribution Functions

While writing my book I had an argument with Feller. He asserted that everyone said "random variable" and I asserted that everyone said "chance variable." We obviously had to use the same name in our books, so we decided the issue by a stochastic procedure. That is, we tossed for it and he won. – Joseph Doob, *Statistical Science*

7.1 Introduction

From the universe of possible information, we ask a question. To address this question, we might collect quantitative data and organize it, for example, using the empirical cumulative distribution function. With this information, we are able to compute sample means, standard deviations, medians and so on.

Similarly, even a fairly simple probability model can have an enormous number of outcomes. For example, flip a coin 333 times. Then the number of outcomes is more than a google (10^{100}) – a number at least 100 quintillion times the number of elementary particles in the known universe. We may not be interested in an analysis that considers separately every possible outcome but rather some simpler concept like the number of heads or the longest run of tails. To focus our attention on the issues of interest, we take a given outcome and compute a number. This function is called a **random variable**.

statistics	**probability**
universe of information	sample space - Ω and probability - P
$\Downarrow$	$\Downarrow$
ask a question and collect data	define a random variable X
$\Downarrow$	$\Downarrow$
organize into the empirical cumulative distribution function	organize into the cumulative distribution function
$\Downarrow$	$\Downarrow$
compute sample means and variances	compute distributional means and variances

Table I: Corresponding notions between statistics and probability. Examining probabilities models and random variables will lead to strategies for the collection of data and inference from these data.

Definition 7.1. *A* **random variable** *is a real valued function from the probability space.*

$$X : \Omega \to \mathbb{R}.$$

Generally speaking, we shall use capital letters near the end of the alphabet, e.g., X, Y, Z for random variables. The range S of a random variable is sometimes called the **state space**.

Exercise 7.2. *Roll a die twice and consider the sample space* $\Omega = \{(i,j); i,j = 1,2,3,4,5,6\}$ *and give some random variables on* Ω.

Exercise 7.3. *Flip a coin 10 times and consider the sample space* Ω*, the set of 10-tuples of heads and tails, and give some random variables on* Ω.

We often create new random variables via composition of functions:

$$\omega \mapsto X(\omega) \mapsto f(X(\omega))$$

Thus, if X is a random variable, then so are

$$X^2, \quad \exp \alpha X, \quad \sqrt{X^2+1}, \quad \tan^2 X, \quad \lfloor X \rfloor$$

and so on. The last of these, rounding down X to the nearest integer, is called the **floor function**.

Exercise 7.4. *How would we use the floor function to round down a number* x *to* n *decimal places.*

7.2 Distribution Functions

Having defined a random variable of interest, X, the question typically becomes, "What are the chances that X lands in some subset of values B?" For example,

$$B = \{\text{odd numbers}\}, \quad B = \{\text{greater than } 1\}, \quad \text{or} \quad B = \{\text{between 2 and 7}\}.$$

We write

$$\{\omega \in \Omega; X(\omega) \in B\} \tag{7.1}$$

to indicate those outcomes ω which have $X(\omega)$, the value of the random variable, in the subset B. We shall often abbreviate (7.1) to the shorter $\{X \in B\}$. Thus, for the example above, we may write the events

$$\{X \text{ is an odd number}\}, \quad \{X \text{ is greater than } 1\} = \{X > 1\}, \quad \{X \text{ is between 2 and 7}\} = \{2 < X < 7\}$$

to correspond to the three choices above for the subset B.

Many of the properties of random variables are not concerned with the specific random variable X given above, but rather depends on the way X *distributes its values*. This leads to a definition in the context of random variables that we saw previously with quantitive data.

Definition 7.5. *A* **(cumulative) distribution function** *of a random variable* X *is defined by*

$$F_X(x) = P\{\omega \in \Omega; X(\omega) \le x\}.$$

Recall that with quantitative observations, we called the analogous notion the *empirical cumulative distribution function*. Using the abbreviated notation above, we shall typically write the less explicit expression

$$F_X(x) = P\{X \le x\}$$

for the distribution function.

Exercise 7.6. *Establish the following identities that relate a random variable the complement of an event and the union and intersection of events*

1. $\{X \in B\}^c = \{X \in B^c\}$
2. *For sets* $B_1, B_2, \ldots$,

$$\bigcup_i \{X \in B_i\} = \{X \in \bigcup_i B\} \quad \textit{and} \quad \bigcap_i \{X \in B_i\} = \{X \in \bigcap_i B\}.$$

3. *If $B_1, \dots B_n$ form a partition of the sample space S, then $C_i = \{X \in B_i\}$, $i = 1, \dots, n$ form a partition of the probability space Ω.*

Exercise 7.7. *For a random variable X and subset B of the sample space S, define*

$$P_X(B) = P\{X \in B\}.$$

Show that P_X is a probability.

For $\{X > x\}$, the complement of $\{X \le x\}$, we have the **survival function**

$$\bar{F}_X(x) = P\{X > x\} = 1 - P\{X \le x\} = 1 - F_X(x).$$

Choose $a < b$, then the event $\{X \le a\} \subset \{X \le b\}$. Their set theoretic difference

$$\{X \le b\} \setminus \{X \le a\} = \{a < X \le b\}.$$

In words, the event that "X is less than or equal to b but not less than or equal to a" is the same event as "X is greater than a and less than or equal to b." Consequently, by the difference rule for probabilities,

$$P\{a < X \le b\} = P(\{X \le b\} \setminus \{X \le a\}) = P\{X \le b\} - P\{X \le a\} = F_X(b) - F_X(a). \quad (7.2)$$

Thus, we can compute the probability that a random variable takes values in an interval by subtracting the distribution function evaluated at the endpoints of the intervals. Care is needed on the issue of the inclusion or exclusion of the endpoints of the interval.

Example 7.8. *To give the cumulative distribution function for X, the sum of the values for two rolls of a die, we start with the table*

x	2	3	4	5	6	7	8	9	10	11	12
$P\{X = x\}$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and create the graph.

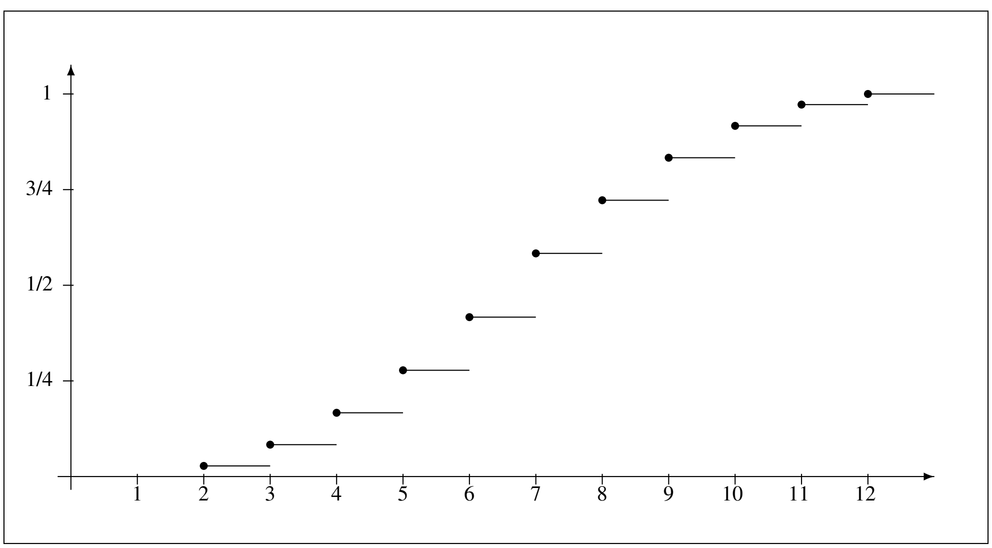

Figure 7.1: Graph of F_X, the cumulative distribution function for the sum of the values for two rolls of a die.

If we look at the graph of this cumulative distribution function, we see that it is constant in between the possible values for X and that the jump size at x is equal to $P\{X = x\}$. In this example, $P\{X = 5\} = 4/36$, the size of the jump at $x = 5$. In addition,

$$F_X(5) - F_X(2) = P\{2 < X \leq 5\} = P\{X = 3\} + P\{X = 4\} + P\{X = 5\} = \sum_{2<x\leq 5} P\{X = x\}$$
$$= \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{9}{36}.$$

We shall call a random variable **discrete** if it has a finite or countably infinite state space. Thus, we have in general that:

$$P\{a < X \leq b\} = \sum_{a<x\leq b} P\{X = x\}.$$

Exercise 7.9. *Let X be the number of heads on three independent flips of a biased coin that turns ups heads with probability p. Give the cumulative distribution function F_X for X.*

Exercise 7.10. *Let X be the number of spades in a collection of three cards. Give the cumulative distribution function for X. Use* R *to plot this function.*

Exercise 7.11. *Find the cumulative distribution function of $Y = X^3$ in terms of F_X, the distribution function for X.*

7.3 Properties of the Distribution Function

A distribution function F_X has the property that it starts at 0, ends at 1 and does not decrease with increasing values of x. This is the content of the next exercise.

Exercise 7.12. *The disribution function F_X has the properties:*

1. $\lim_{x\to-\infty} F_X(x) = 0$.
2. $\lim_{x\to\infty} F_X(x) = 1$.
3. *F_X is nondecreasing.*

7.3.1 Discrete Random Variables

The cumulative distribution function F_X of a discrete random variable X is constant except for jumps. At the jump, F_X is **right continuous**,

$$\lim_{x\to x_0+} F_X(x) = F_X(x_0). \qquad (7.3)$$

The next exercise ask that this be shown more generally.

Exercise 7.13. *Prove the statement (7.3) concerning the right continuity of the distribution function from the continuity property of a probability.*

Exercise 7.14. *Show that for any x_0,*

$$P\{X < x_0\} = \lim_{x\to x-} F_X(x) = F_X(x_0-),$$

the left limit of F_X at x_0.

Putting the previous two exercises together, we find that

$$P\{X = x_0\} = P(\{X \leq x_0\} \setminus \{X < x_0\}) = P\{X \leq x_0\} - P\{X < x_0\} = F_X(x_0) - F_X(x_0-),$$

The size of the jump in $F_X(x)$ at the value x_0.

7.3.2 Continuous Random Variables

Definition 7.15. *A* **continuous random variable** *has a cumulative distribution function F_X that is differentiable.*

So, distribution functions for continuous random variables increase smoothly. To show how this can occur, we will develop an example of a continuous random variable.

Example 7.16. *Consider a dartboard having unit radius. Assume that the dart lands randomly uniformly on the dartboard.*

Let X be the distance from the center. For $x \in [0,1]$,

$$F_X(x) = P\{X \leq x\} = \frac{\text{area inside circle of radius } x}{\text{area of circle}} = \frac{\pi x^2}{\pi 1^2} = x^2.$$

Thus, we have the distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x^2 & \text{if } 0 < x \leq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

The first line states that X cannot be negative. The third states that X is at most 1, and the middle lines describes how X distributes is values between 0 and 1. For example,

$$F_X\left(\frac{1}{2}\right) = \frac{1}{4}$$

indicates that with probability 1/4, the dart will land within 1/2 unit of the center of the dartboard.

Exercise 7.17. *Find the probability that the dart lands between 1/3 unit and 2/3 unit from the center. Find the median, the first quartile, and the third quartiles.*

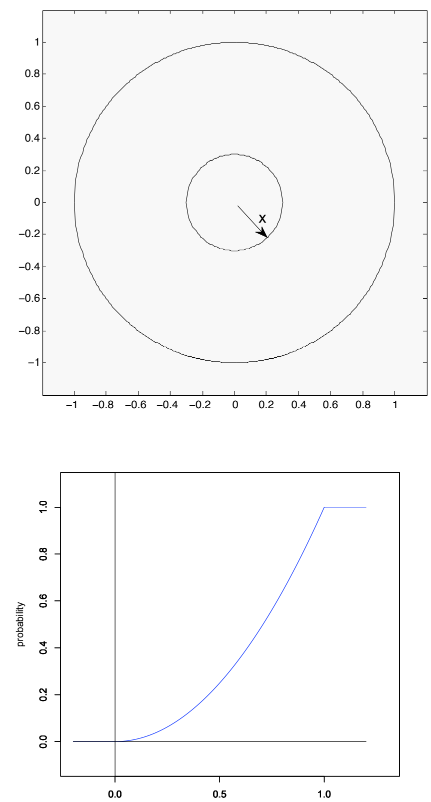

Figure 7.2: (top) Dartboard. (bottom) Cumulative distribution function for the dartboard random variable.

Exercise 7.18. *Let the reward Y for throwing the dart be the inverse $1/X$ of the distance from the center. Find the cumulative distribution function for Y.*

Exercise 7.19. *An* **exponential random variable** *X has cumulative distribution function*

$$F_X(x) = P\{X \leq x\} = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - \exp(-\lambda x) & \text{if } x > 0. \end{cases} \tag{7.4}$$

for some $\lambda > 0$. Show that F_X has the properties of a distribution function.

We can create an expression and perform an evaluation using R.

```
> F<-expression(1-exp(-lambda*x))
```

We can then evaluate $F_X(3)$ and $F_X(1)$ with $\lambda = 2$ as follows.

```
> x<-c(10,30);lambda<-1/10
> (Feval<-eval(F))
[1] 0.6321206 0.9502129
> Feval[2]-Feval[1]
[1] 0.3180924
```

The last expression gives the value for $F_X(30) - F_X(10) = P\{10 < X \leq 30\}$.

This function is also stored in R and so its value at x can be computed in R using the command `pexp(x,0.1)` for $\lambda = 1/10$. Thus, we make the computation above by

```
> pexp(30,0.1)-pexp(10,0.1)
[1] 0.3180924
```

We can draw the distribution function using the `curve` command.

```
> curve(pexp(x,0.1),0,80)
```

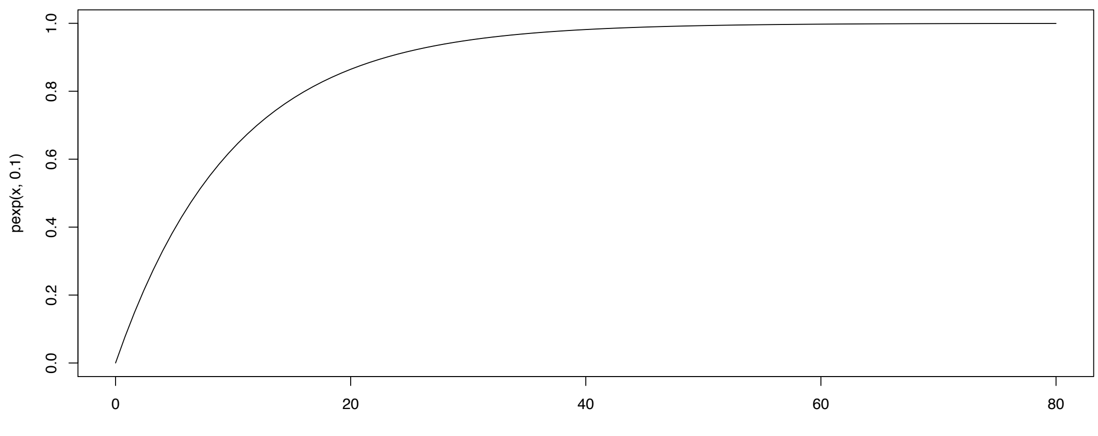

Figure 7.3: Cumulative distribution function for an exponential random variable with $\lambda = 1/10$.

Exercise 7.20. *The time until the next bus arrives is an exponential random variable with* $\lambda = 1/10$ *minutes. A person waits at the bus stop until the bus arrives, giving up when the wait reaches 20 minutes. Give the cumulative distribution function for* T*, the time that the person remains at the bus station and sketch a graph.*

Even though the cumulative distribution function is defined for every random variable, we will often use other characterizations, namely, the **mass function** for discrete random variable and the **density function** for continuous random variables. Indeed, we typically will introduce a random variable via one of these two functions. In the next two sections we introduce these two concepts and develop some of their properties.

7.4 Mass Functions

Definition 7.21. *The* **(probability) mass function** *of a discrete random variable* X *is*

$$f_X(x) = P\{X = x\}.$$

The mass function has a value at x equal to the size of the jump in the distribution function. In symbols,

$$f_X(x) = F_X(x) - F_X(x-)$$

where $F_X(x-)$ is the left limit of F_X at x.

The mass function has two basic properties:

- $f_X(x) \geq 0$ for all x in the state space.
- $\sum_x f_X(x) = 1$.

The first property is based on the fact that probabilities are non-negative. The second follows from the observation that the collection $C_x = \{\omega; X(\omega) = x\}$ for all $x \in S$, the state space for X, forms a partition of the probability space Ω. In Example 7.8, we saw the mass function for the random variable X that is the sum of the values on two independent rolls of a fair dice.

Example 7.22. *Let's make tosses of a biased coin whose outcomes are independent. We shall continue tossing until we obtain a toss of heads. Let X denote the random variable that gives the number of tails before the first head and p denote the probability of heads in any given toss. Then*

$$\begin{aligned}
f_X(0) &= P\{X = 0\} = P\{H\} = p \\
f_X(1) &= P\{X = 1\} = P\{TH\} = (1-p)p \\
f_X(2) &= P\{X = 2\} = P\{TTH\} = (1-p)^2 p \\
\vdots \quad & \qquad \vdots \qquad\qquad \vdots \\
f_X(x) &= P\{X = x\} = P\{T\cdots TH\} = (1-p)^x p
\end{aligned}$$

So, the probability mass function $f_X(x) = (1-p)^x p$. Because the terms in this mass function form a geometric sequence, X is called a **geometric random variable***. Recall that a geometric sequence $c, cr, cr^2, \ldots, cr^n$ has sum*

$$s_n = c + cr + cr^2 + \cdots + cr^n = \frac{c(1-r^{n+1})}{1-r}$$

for $r \neq 1$. If $|r| < 1$, then $\lim_{n\to\infty} r^n = 0$ and thus s_n has a limit as $n \to \infty$. In this case, the infinite sum is the limit

$$c + cr + cr^2 + \cdots + cr^n + \cdots = \lim_{n\to\infty} s_n = \frac{c}{1-r}. \tag{7.5}$$

Exercise 7.23. *Establish the formula (7.5) above for s_n.*

The mass function above forms a geometric sequence with the ratio $r = 1-p$. Consequently, for positive integers a and b,

$$\begin{aligned}
P\{a < X \leq b\} &= \sum_{x=a+1}^{b} (1-p)^x p = (1-p)^{a+1}p + \cdots + (1-p)^b p \\
&= \frac{(1-p)^{a+1}p - (1-p)^{b+1}p}{1-(1-p)} = (1-p)^{a+1} - (1-p)^{b+1}
\end{aligned}$$

We can take $a = -1$ to find the distribution function for a geometric random variable.

$$F_X(b) = P\{X \leq b\} = 1 - (1-p)^{b+1}. \tag{7.6}$$

To obtain (7.6) in another way, note that the event $\{X \geq b+1\} = \{X > b\}$ is the same as having the first $b+1$ coin tosses turn up *tails*. This event consists of $b+1$ independent events each with probability $1-p$. Thus, $P\{X \geq b+1\} = P\{X > b\} = (1-p)^{b+1}$. By noting that the distribution function, $F_X(b) = 1 - P\{X > b\}$, we again obtain (7.6).

Exercise 7.24. *Show that for a geometric random variable X,*

$$P\{X \geq a + b | X \geq b\} = P\{X \geq a\}. \tag{7.7}$$

This property is called **memorylessness***. In words, if the first b trials results in failures, then the probability of at least a additional failures is the same as the probability of at least a failures from the beginning. The fact that we begin with b failures does not impact the number of trials afterwards until a success.*

Conversely, if the memoryless property holds for an $\mathbb{N}$-valued random variable X, then X is a geometric random variable.

The mass function and the cumulative distribution function for the geometric random variable with parameter $p = 1/3$ can be found in R by writing

```
> x<-0:10
> f<-dgeom(x,1/3)
> F<-pgeom(x,1/3)
```

The initial `d` indicates **d**ensity and `p` indicates the **p**robability from the distribution function.

```
> data.frame(x,f,F)
    x           f         F
1   0 0.333333333 0.3333333
2   1 0.222222222 0.5555556
3   2 0.148148148 0.7037037
4   3 0.098765432 0.8024691
5   4 0.065843621 0.8683128
6   5 0.043895748 0.9122085
7   6 0.029263832 0.9414723
8   7 0.019509221 0.9609816
9   8 0.013006147 0.9739877
10  9 0.008670765 0.9826585
11 10 0.005780510 0.9884390
```

Note that the difference in values in the distribution function $F_X(x) - F_X(x-1)$, giving the height of the jump in F_X at x, is equal to the value of the mass function. For example,

$$F_X(3) - F_X(2) = 0.8024691 - 0.7037037 = 0.0987654 = f_X(3).$$

Exercise 7.25. *Check that the jumps in the cumulative distribution function for the geometric random variable above is equal to the values of the mass function.*

Exercise 7.26. *For the geometric random variable above, find* $P\{X \le 3\}$, $P\{2 < X \le 5\}$. $P\{X > 4\}$.

We can simulate 100 geometric random variables with parameter $p = 1/3$ using the R command `rgeom(100,1/3)`. (See Figure 7.4.)

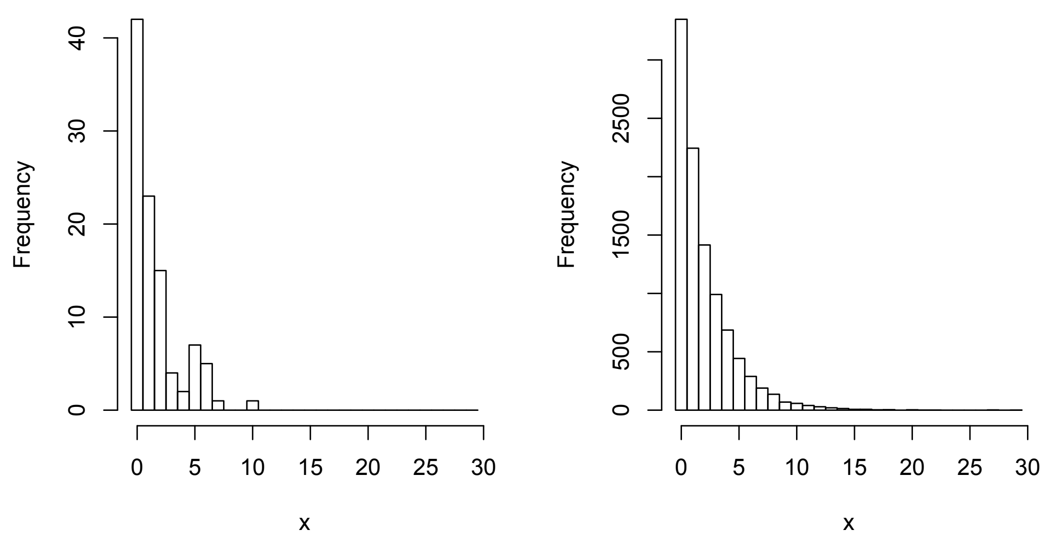

Figure 7.4: Histogram of 100 and 10,000 simulated geometric random variables with $p = 1/3$. Note that the histogram looks much more like a geometric series for 10,000 simulations. We shall see later how this relates to the law of large numbers.

7.5 Density Functions

Definition 7.27. *For X a random variable whose distribution function F_X has a derivative. The function f_X satisfying*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

is called the **probability density function** *and X is called a* **continuous random variable**.

By the fundamental theorem of calculus, the density function is the derivative of the distribution function.

$$f_X(x) = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = F_X'(x).$$

In other words,

$$F_X(x + \Delta x) - F_X(x) \approx f_X(x)\Delta x.$$

We can compute probabilities by evaluating definite integrals

$$P\{a < X \le b\} = F_X(b) - F_X(a) = \int_a^b f_X(t)\,dt.$$

The density function has two basic properties that mirror the properties of the mass function:

- $f_X(x) \ge 0$ for all x in the state space.
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

Return to the dart board example, letting X be the distance from the center of a dartboard having unit radius. Then,

$$\begin{aligned} P\{x < X \le x + \Delta x\} &= F_X(x + \Delta x) - F_X(x) \\ &\approx f_X(x)\Delta x = 2x\Delta x \end{aligned}$$

and X has density

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 2x & \text{if } 0 \le x \le 1, \\ 0 & \text{if } x > 1. \end{cases}$$

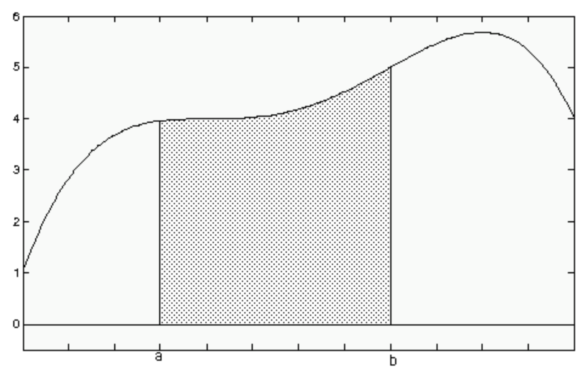

Figure 7.5: The probability $P\{a < X \le b\}$ is the area under the density function, above the x axis between $y = a$ and $y = b$.

Exercise 7.28. *Let f_X be the density for a random variable X and pick a number x_0. Explain why $P\{X = x_0\} = 0$.*

Exercise 7.29. *Plot, on both the distribution function and the density function, the probability that the dart lands between 1/3 unit and 2/3 unit from the center.*

Example 7.30. *For the exponential distribution function (7.4), we have the density function*

$$f_X(x) = \begin{cases} 0 & \text{if } x \le 0, \\ \lambda e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

R *performs differentiation. We must first create an expression*

```
> F<-expression(1-exp(-lambda*x))
```

We then differentiate using the `D` *command, placing x, the variable of differentiation in quotes.*

```
> f<-D(F,"x")
> f
exp(-lambda * x) * lambda
```

Example 7.31. *Density functions do not need to be bounded, for example, if we take*

$$f_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{c}{\sqrt{x}} & \text{if } 0 < x < 1, \\ 0 & \text{if } 1 \leq x. \end{cases}$$

Then, to find the value of the constant c, we compute the integral

$$1 = \int_0^1 \frac{c}{\sqrt{t}}\, dt = 2c\sqrt{t}\Big|_0^1 = 2c.$$

So $c = 1/2$. *For* $0 \leq a < b \leq 1$,

$$P\{a < X \leq b\} = \int_a^b \frac{1}{2\sqrt{t}}\, dt = \sqrt{t}\Big|_a^b = \sqrt{b} - \sqrt{a}.$$

Exercise 7.32. *Give the cumulative distribution function for the random variable in the previous example.*

Exercise 7.33. *Let X be a continuous random variable with density f_X, then the random variable $Y = aX + b$ has density*

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

(Hint: Begin with the definition of the cumulative distribution function F_Y for Y. Consider the cases $a > 0$ and $a < 0$ separately.)

7.6 Mixtures

Exercise 7.34. *Let F_1 and F_2 be two cumulative distribution functions and let $\pi \in (0, 1)$, then*

$$F(x) = \pi F_1(x) + (1 - \pi)F_2(x)$$

is a cumulative distribution function.

We call the distribution F a **mixture** of F_1 and F_2. Mixture distributions occur routinely. To see this, first flip a coin, heads occurring with probability π. In this case the random variable

$$X = \begin{cases} X_1 & \text{if the coin lands heads,} \\ X_2 & \text{if the coin lands tails.} \end{cases}$$

If X_i has distribution function F_i, $i = 1, 2$, then, by the law of total probability,

$$\begin{aligned} F_X(x) = P\{X \leq x\} &= P\{X \leq x|\text{coin lands heads}\}P\{\text{coin lands heads}\} \\ &\quad + P\{X \leq x|\text{coin lands tails}\}P\{\text{coin lands tails}\} \\ &= P\{X_1 \leq x\}\pi + P\{X_2 \leq x\}(1-\pi) = \pi F_1(x) + (1-\pi)F_2(x) \end{aligned}$$

More generally, let $X_1, \ldots, X_n$ be random variables with distribution functions $F_1, \ldots, F_n$ and $\pi_1, \ldots, \pi_n$ be positive numbers with $\sum_{i=1}^n \pi_i = 1$. In this case, roll an n sided die, i showing with probability π_i. If the die shows i, then we use the random variable X_i. To be concrete, individuals arriving to take an airline flight are assigned to group i with probability π_i. Let X_i be the (random) time until individuals in group i are seated. Then the distribution function for the time to be seated

$$\begin{aligned} F_X(x) = P\{X \le x\} &= \sum_{i=1}^n P\{X \le x | \text{assigned group } i\} P\{\text{assigned group } i\} \\ &= \sum_{i=1}^n P\{X_i \le x\}\pi_i = \pi_1 F_1(x) + \cdots + \pi_n F_n(x). \end{aligned}$$

F is call the **mixture** of $F_1, \ldots, F_n$ with weights $\pi_1, \ldots, \pi_n$.

If the X_i are discrete random variables, then so is X. The **mass function** for X is

$$\begin{aligned} f_X(x) = F_X(x) - F_X(x-) &= \pi_1(F_1(x) - F_1(x-)) + \cdots + \pi_n(F_n(x) - F_n(x-)) \\ &= \pi_1 f_1(x) + \cdots + \pi_n f_n(x). \end{aligned}$$

Exercise 7.35. *Check that f_X is a mass function.*

Exercise 7.36. *Find the mass function for the mixture of the three mass functions*

x	$f_1(x)$	$f_2(x)$	$f_3(x)$
1	0.2	0.5	0.1
2	0.3	0.5	0.1
3	0.1	0	0.2
4	0.4	0	0.2
5	0	0	0.4

and weights $\pi = (1/4, 1/4, 1/2)$,

If the X_i are continuous random variables, then so is X. The **density function** for X is

$$\begin{aligned} f_X(x) = F_X'(x) &= \pi_1 F_1'(x) + \cdots + \pi_n F_n'(x) \\ &= \pi_1 f_1(x) + \cdots + \pi_n f_n(x) \\ &= \sum_{i=1}^n f_i(x)\pi_i. \end{aligned}$$

Checking that f_X is a density function is similar to the exercise above. Just replace the sum on x with an integral.

7.7 Joint and Conditional Distributions

Because we will collect data on several observations, we must, as well, consider more than one random variable at a time in order to model our experimental procedures. Consequently, we will expand on the concepts above to the case of multiple random variables and their joint distribution. For the case of two random variables, X_1 and X_2, this means looking at the probability of events,

$$P\{X_1 \in B_1, X_2 \in B_2\}.$$

For discrete random variables, take $B_1 = \{x_1\}$ and $B_2 = \{x_2\}$. Then, we have

7.7.1 Discrete Random Variables

Definition 7.37. *The* **joint probability mass function**

$$f_{X_1,X_2}(x_1, x_2) = P\{X_1 = x_1, X_2 = x_2\}.$$

The mass functions for X_1 and X_2 can be obtained from the joint mass function by summing over the values for the other random variable. Thus, for example,

$$f_{X_1}(x_1) = P\{X_1 = x_1\} = \sum_{x_2} P\{X_1 = x_1, X_2 = x_2\} = \sum_{x_2} f_{X_1,X_2}(x_1, x_2). \tag{7.8}$$

In this case, we use the expression **marginal probability mass function** to distinguish it from the joint probability mass function.

Exercise 7.38. *Let X_1 and X_2 have the joint mass function displayed in the table below*

	$f_{X_1,X_2}(x_1,x_2)$					
$x_2\backslash x_1$	1	2	3	4	5	
-1	0.09	0.04	0.03	0.01	0.02	
0	0.07	0	0.07	0.02	0.03	
1	0.10	0.06	0.05	0.08	0.06	
2	0.01	0.08	0.09	0.05	0.04	

Show that the sum of the entries is 1 and determine the marginal mass functions.

The **conditional mass functions** looks at the probabilities that one random variable takes on a given value, given a value for the second random variable. The conditional mass function of X_2 given X_1 is denoted $f_{X_2|X_1}(x_2|x_1) = P\{X_2 = x_2|X_1 = x_1\}$. To compute this function,

$$f_{X_2|X_1}(x_2|x_1) = P\{X_2 = x_2|X_1 = x_1\} = \frac{P\{X_1 = x_1, X_2 = x_2\}}{P\{X_1 = x_1\}} = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \tag{7.9}$$

provided $f_{X_1}(x_1) > 0$.

Exercise 7.39. *Show that, for each value of x_1, $f_{X_2|X_1}(x_2|x_1)$ is a mass function, that is, the values are non-negative and the sum over all values for x_2 equals 1.*

Exercise 7.40. *For each value of x_1, find the conditional mass function. $f_{X_2|X_1}(x_2|x_1)$ for the values in the table above.*

7.7.2 Continuous Random Variables

For continuous random variables, we consider $B_1 = (x_1, x_1 + \Delta x_1]$ and $B_2 = (x_2, x_2 + \Delta x_2]$ and ask that for some function f_{X_1,X_2}, the **joint probability density function** to satisfy

$$P\{x_1 < X_1 \le x_1 + \Delta x_1, x_2 < X_2 \le x_2 + \Delta x_2\} \approx f_{X_1,X_2}(x_1,x_2)\Delta x_1 \Delta x_2.$$

Similar to mass functions, the density functions for X_1 and X_2 can be obtained from the joint density function by integrating over the values for the other random variable. Also, we sometimes say **marginal probability density function** to distinguish it from the joint probability density function. Thus, for example, in analogy with (7.8).

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\,dx_2. \tag{7.10}$$

We can obtain this identity starting with (7.8) and using Riemann sums in a manner similar to the argument that led to the formula for expectation for a continuous random variable.

For the **conditional density**, we start with

$$\begin{aligned} P\{x_2 < X_2 \le x_2 + \Delta x_2 | x_1 < X_1 \le x_1 + \Delta x_1\} &= \frac{P\{x_1 < X_1 \le x_1 + \Delta x_1, x_2 < X_2 \le x_2 + \Delta x_2\}}{P\{x_1 < X_1 \le x_1 + \Delta x_1\}} \\ &\approx \frac{f_{X_1,X_2}(x_1,x_2)\Delta x_1 \Delta x_2}{f_{X_1}(x_1)\Delta x_1} = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}\Delta x_2 \end{aligned}$$

Next, divide by Δx_2 and let $\Delta x_2 \to 0$. In keeping with the analogies between discrete and continuous densities, we have the following definition.

Definition 7.41. *The conditional density function*

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}$$

provided $f_{X_1}(x_1) > 0$.

Exercise 7.42. *Show that, for each value of* x_1*,* $f_{X_2|X_1}(x_2|x_1)$ *is a density function, that is, the values are non-negative and the integral over all values for* x_2 *equals 1.*

Exercise 7.43. *Verify that*

$$f_{X_1,X_2}(x_1,x_2) \begin{cases} x_1 + \frac{3}{2}x_2^2 & 0 < x_1 \le 1, 0 < x_2 \le 1. \\ 0 & \text{otherwise} \end{cases}$$

is a joint density function. Find the marginal densities.

7.7.3 Independent Random Variables

Many of our experimental protocols will be designed so that observations are independent. More precisely, we will say that two random variables X_1 and X_2 are **independent** if any two events associated to them are independent, i.e.,

$$P\{X_1 \in B_1, X_2 \in B_2\} = P\{X_1 \in B_1\}P\{X_2 \in B_2\}.$$

In words, the probability that the two events $\{X_1 \in B_1\}$ and $\{X_2 \in B_2\}$ happen simultaneously is equal to the product of the probabilities that each of them happen individually.

For independent discrete random variables, we have that

$$f_{X_1,X_2}(x_1,x_2) = P\{X_1 = x_1, X_2 = x_2\} = P\{X_1 = x_1\}P\{X_2 = x_2\} = f_{X_1}(x_1)f_{X_2}(x_2).$$

In this case, we say that the joint probability mass function is the product of the **marginal mass functions**.

For continuous random variables,

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2)\Delta x_1 \Delta x_2 &\approx P\{x_1 < X_1 \le x_1 + \Delta x_1, x_2 < X_2 \le x_2 + \Delta x_2\} \\ &= P\{x_1 < X_1 \le x_1 + \Delta x_1\}P\{x_2 < X_2 \le x_2 + \Delta x_2\} \approx f_{X_1}(x_1)\Delta x_1 f_{X_2}(x_2)\Delta x_2 \\ &= f_{X_1}(x_1)f_{X_2}(x_2)\Delta x_1 \Delta x_2. \end{aligned}$$

Thus, for independent continuous random variables, the joint probability density function

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

is the product of the **marginal density functions**.

Exercise 7.44. *Generalize the notion of independent mass and density functions to more than two random variables.*

Soon, we will be looking at n independent observations $x_1, x_2, \ldots, x_n$ arising from an unknown density or mass function f. Thus, the joint density is

$$f(x_1)f(x_2)\cdots f(x_n).$$

Generally speaking, the density function f will depend on the choice of a parameter value θ. (For example, the unknown parameter in the density function for an exponential random variable that describes the waiting time for a bus.) Given the data from the n observations, the **likelihood function** arises by considering this joint density not as a function of $x_1, \ldots, x_n$, but rather as a function of the parameter θ. We shall learn how the study of the likelihood plays a major role in parameter estimation and in the testing of hypotheses.

7.8 Simulating Random Variables

One goal for these notes is to provide the tools needed to design inferential procedures based on sound principles of statistical science. Thus, one of the very important uses of statistical software is the ability to generate pseudo-data to simulate the actual data. This provides the opportunity to explore the properties of the data through simulation and to test and refine methods of analysis in advance of the need to use these methods on genuine data. For many of the frequently used families of random variables, R provides commands for their simulation. We shall examine these families and their properties in Topic 9, *Examples of Mass Functions and Densities*. For other circumstances, we will need to have methods for simulating sequence of independent random variables that possess a common distribution. We first consider the case of discrete random variables.

7.8.1 Discrete Random Variables and the `sample` Command

The `sample` command is used to create simple and stratified random samples. Thus, if we enter a sequence `x`, `sample(x,40)` chooses 40 entries from `x` in such a way that all choices of size 40 have the same probability.

This uses the default R command of **sampling without replacement**. We can use this command to simulate discrete random variables. To do this, we need to give the state space in a vector `x` and a mass function `f`. The call for `replace=TRUE` indicates that we are **sampling with replacement**. Then to give a sample of n independent random variables having common mass function `f`, we use `sample(x,n,replace=TRUE,prob=f)`.

Example 7.45. *Let X be described by the mass function*

x	1	2	3	4
$f_X(x)$	0.1	0.2	0.3	0.4

Then to simulate 50 independent observations from this mass function:

```
> x<-c(1,2,3,4); f<-c(0.1,0.2,0.3,0.4)
> sum(f)
[1] 1
> data<-sample(x,50,replace=TRUE,prob=f)
> data
  [1] 1 4 4 4 4 4 3 3 4 3 3 2 3 3 3 4 4 3 3 2 4 1 3 3 4 2 3 3 3 1 2 4 3 2 3 4 4 4 4 2 4 1
[43] 2 3 4 4 1 4 3 4
```

Notice that 1 is the least represented value and 4 is the most represented. If the command `prob=f` *is omitted, then* `sample` *will choose uniformly from the values in the vector* `x`*. Let's check our simulation against the mass function that generated the data. First, recount the observations that take on each possible value for x. We can make a table.*

```
> table(data)
data
 1  2  3  4
 5  7 18 20
```

or use the counts to determine the simulated proportions.

```
> counts<-numeric(4)
> for (i in 1:4){counts[i]<-sum(data==i)}
> simprob<-counts/(sum(counts))
> data.frame(x,f,simprob)
  x   f simprob
1 1 0.1 0.10
2 2 0.2 0.14
3 3 0.3 0.36
4 4 0.4 0.40
```

The expression `data==i` *returns a sequence* `FALSE` *and* `TRUE`*. the* `sum` *command adds up the number of times* `TRUE` *appears.*

Exercise 7.46. *Simulate the sums on each of 20 rolls of a pair of dice. Repeat this for 1000 rolls and compare the simulation with the appropriate mass function.*

Exercise 7.47. *Simulate the mixture in Exercise 7.36 and comment on how it matches the mixture mass function.*

7.8.2 Continuous Random Variables and the Probability Transform

If X a continuous random variable with a density f_X that is positive everywhere in its domain, then the distribution function $F_X(x) = P\{X \le x\}$ is strictly increasing. In this case F_X has a inverse function F_X^{-1}, known as the **quantile function**.

Exercise 7.48. $F_X(x) \le u$ *if and only if* $x \le F_X^{-1}(u)$.

The **probability transform** follows from an analysis of the random variable

$$U = F_X(X)$$

Note that F_X has range from 0 to 1. It cannot take values below 0 or above 1. Thus, U takes on values between 0 and 1 and, therefore,

$$F_U(u) = 0 \text{ for } u < 0 \quad \text{and} \quad F_U(u) = 1 \text{ for } u \ge 1.$$

For values of u between 0 and 1, note that

$$P\{F_X(X) \le u\} = P\{X \le F_X^{-1}(u)\} = F_X(F_X^{-1}(u)) = u.$$

Taken together, we have the distribution function for the random variable U,

$$F_U(u) = \begin{cases} 0 & u < 0, \\ u & 0 \le u < 1, \\ 1 & 1 \le u. \end{cases}$$

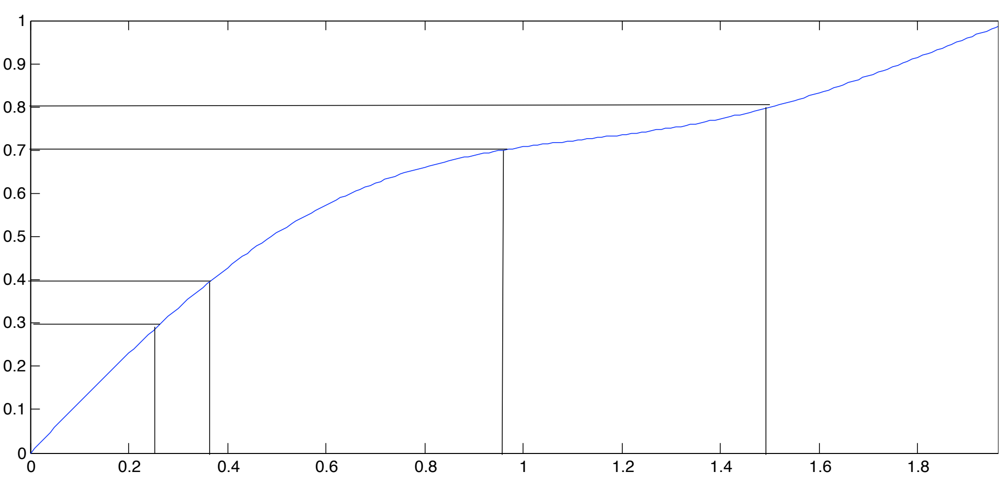

Figure 7.6: Illustrating the Probability Transform. First simulate uniform random variables $u_1, u_2, \ldots, u_n$ on the interval $[0, 1]$. About 10% of the random numbers should be in the interval $[0.3, 0.4]$. This corresponds to the 10% of the simulations on the interval $[0.28, 0.38]$ for a random variable with distribution function F_X shown. Similarly, about 10% of the random numbers should be in the interval $[0.7, 0.8]$ which corresponds to the 10% of the simulations on the interval $[0.96, 1.51]$ for a random variable with distribution function F_X, These values on the x-axis can be obtained from taking the inverse function of F_X, i.e., $x_i = F_X^{-1}(u_i)$.

If we can simulate U, we can simulate a random variable with distribution F_X via the quantile function

$$X = F_X^{-1}(U). \tag{7.11}$$

Take a derivative of $F_U(u)$ to see that its density

$$f_U(u) = \begin{cases} 0 & u < 0, \\ 1 & 0 \le u < 1, \\ 0 & 1 \le u. \end{cases}$$

Because the random variable U has a constant density over the interval of its possible values, it is called **uniform** on the interval $[0, 1]$. It is simulated in R using the `runif` command. The identity (7.11) is called the **probability transform**. This transform is illustrated in Figure 7.6. We can see how the probability transform works in the following example.

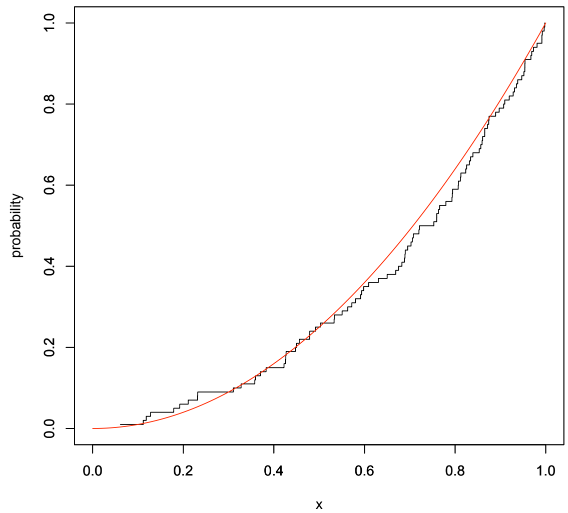

Figure 7.7: The distribution function (red) and the empirical cumulative distribution function (black) based on 100 simulations of the dart board distribution. R commands given below.

Example 7.49. *For the dart board, for x between 0 and 1, the distribution function $u = F_X(x) = x^2$ and thus the quantile function*

$$x = F_X^{-1}(u) = \sqrt{u}.$$

We can simulate independent observations of the distance from the center $X_1, X_2, \ldots, X_n$ of the dart board by simulating independent uniform random variables $U_1, U_2, \ldots U_n$ and taking the quantile function

$$X_i = \sqrt{U_i}.$$

```
> u<-runif(100)
> xu<-sqrt(u)
> plot(sort(xu),1:length(xu)/length(xu),
+   type="s",xlim=c(0,1),ylim=c(0,1), xlab="x",ylab="probability")
> x<-seq(0,1,0.01)
> lines(x,x^2,col="red") #add the distribution function to the graph
```

We have used the `lines` *command to ad the distribution function $F_X(x) = x^2$. Notice how it follows the empirical cumulative distribution function.*

Exercise 7.50. *If U is uniform on $[0, 1]$, then so is $V = 1 - U$.*

Sometimes, it is easier to simulate X using $F_X^{-1}(V)$.

Example 7.51. *For an exponential random variable, set*

$$u = F_X(x) = 1 - \exp(-\lambda x), \text{ and thus } x = -\frac{1}{\lambda}\ln(1-u)$$

Consequently, we can simulate independent exponential random variables $X_1, X_2, \ldots, X_n$ by simulating independent uniform random variables $V_1, V_2, \ldots V_n$ and taking the transform

$$X_i = -\frac{1}{\lambda}\ln V_i.$$

R *accomplishes this directly through the* `rexp` *command.*