4

Conditional Probability and Expectation

4.1 INTRODUCTION

We have thus far defined the conditional probability $P(B \mid A)$ only when $P(A) > 0$. However, there are many situations when it is natural to talk about a conditional probability given an event of probability 0. For example, suppose that a real number R is selected at random, with density f. If R takes the value x, a coin with probability of heads $g(x)$ is tossed ($0 \leq g(x) \leq 1$). It is natural to assert that the conditional probability of obtaining a head, given $R = x$, is $g(x)$. But since R is absolutely continuous, the event $\{R = x\}$ has probability 0, and thus conditional probabilities given $R = x$ are not as yet defined.

If we ignore this problem for the moment, we can find the over-all probability of obtaining a head by the following intuitive argument. The probability that R will fall into the interval $(x, x + dx]$ is roughly $f(x)\,dx$; given that R falls into this interval, the probability of a head is roughly $g(x)$. Thus we should expect, from the theorem of total probability, that the probability of a head will be $\sum_x g(x)f(x)\,dx$, which approximates $\int_{-\infty}^{\infty} g(x)f(x)\,dx$. Thus the probability in question is a weighted average of conditional probabilities, the weights being assigned in accordance with the density f.

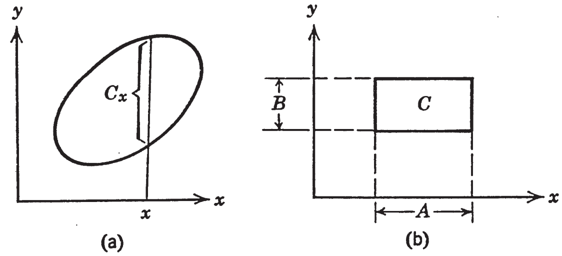

FIGURE 4.1.1

Let us examine what is happening here. We have two random variables R_1 and R_2 [$R_1 = R$, $R_2 =$ (say) the number of heads obtained]. We are specifying the density of R_1, and for each x and each Borel set B we are specifying a quantity $P_x(B)$ that is to be interpreted intuitively as the conditional probability that $R_2 \in B$ given that $R_1 = x$. (We shall often write $P\{R_2 \in B \mid R_1 = x\}$ for $P_x(B)$.)

We would like to conclude that the probabilities of all events involving R_1 and R_2 are now determined. Suppose that C is a two-dimensional Borel set. What is a reasonable figure for $P\{(R_1, R_2) \in C\}$? Intuitively, the probability that R_1 falls into $(x, x + dx]$ is $f_1(x)\,dx$. Given that this happens, that is, (roughly) given $R_1 = x$, the only way (R_1, R_2) can lie in C is if R_2 belongs to the "section" $C_x = \{y\colon (x, y) \in C\}$ (see Figure 4.1.1a). This happens with probability $P_x(C_x)$. Thus we expect that the total probability that (R_1, R_2) will belong to C is

$$\int_{-\infty}^{\infty} P_x(C_x) f_1(x)\,dx$$

In particular, if $C = A \times B = \{(x, y)\colon x \in A,\ y \in B\}$ (see Figure 4.1.1b),

$$C_x = \varnothing \qquad \text{if } x \notin A; \qquad C_x = B \qquad \text{if } x \in A$$

Thus

$$P\{(R_1, R_2) \in C\} = P\{R_1 \in A, R_2 \in B\} = \int_A P_x(B) f_1(x)\,dx$$

The above reasoning may be formalized as follows. Let $\Omega = E^2$, $\mathscr{F} =$ Borel subsets, $R_1(x, y) = x$, $R_2(x, y) = y$. Let f_1 be a density function on E^1, that is, a nonnegative function such that $\int_{-\infty}^{\infty} f_1(x)\,dx = 1$. Suppose that for each real x we are given a probability measure P_x on the Borel subsets of E^1. Assume also that $P_x(B)$ is a piecewise continuous function of x for each fixed B.

Then it turns out that there is a unique probability measure P on $\mathscr{F}$ such that for all Borel subsets A, B of E^1

$$P(A \times B) = \int_A P_x(B) f_1(x)\, dx \tag{4.1.1}$$

Thus the requirement (4.1.1), which may be regarded as a continuous version of the theorem of total probability, determines P uniquely. In fact, if $C \in \mathscr{F}$, $P(C)$ is given explicitly by

$$P(C) = \int_{-\infty}^{\infty} P_x(C_x) f_1(x)\, dx \tag{4.1.2}$$

Notice that if $R_1(x, y) = x$, $R_2(x, y) = y$, then

$$P(A \times B) = P\{R_1 \in A,\ R_2 \in B\}$$

and

$$P(C) = P\{(R_1, R_2) \in C\}$$

Furthermore, the distribution function of R_1 is given by

$$\begin{aligned} F_1(x_0) = P\{R_1 \le x_0\} &= P\{R_1 \in A,\ R_2 \in B\} \\ &\qquad \text{where } A = (-\infty, x_0],\ B = (-\infty, \infty) \\ &= \int_A P_x(B) f_1(x)\, dx = \int_{-\infty}^{x_0} f_1(x)\, dx \end{aligned}$$

Thus f_1 is in fact the density of R_1. Notice also that

$$P\{R_2 \in B\} = P\{R_1 \in A,\ R_2 \in B\}$$

where $A = (-\infty, \infty)$; hence

$$P\{R_2 \in B\} = \int_{-\infty}^{\infty} P_x(B) f_1(x)\, dx \tag{4.1.3}$$

To summarize: If we start with a density for R_1 and a set of probabilities $P_x(B)$ that we interpret as $P\{R_2 \in B \mid R_1 = x\}$, the probabilities of events of the form $\{(R_1, R_2) \in C\}$ are determined in a natural way, if you believe that there should be a continuous version of the theorem of total probability; $P\{(R_1, R_2) \in C\}$ is given explicitly by (4.1.2), which reduces to (4.1.1) in the special case when $C = A \times B$.

We have not yet answered the question of how to define $P\{R_2 \in B \mid R_1 = x\}$ for arbitrarily specified random variables R_1 and R_2; we attack this problem later in the chapter. Instead we have approached the problem in a somewhat oblique way. However, there are many situations in which one specifies the density of R_1, and then the conditional probability of events involving R_2, given $R_1 = x$. We now know how to formulate such problems precisely. Consider again the problem at the beginning of the section. If R_1 has density f, and a coin with probability of heads $g(x)$ is tossed whenever $R_1 = x$ (and a head corresponds to $R_2 = 1$, a tail to $R_2 = 0$), then the probability of obtaining a head is

$$P\{R_2 = 1\} = \int_{-\infty}^{\infty} P\{R_2 = 1 \mid R_1 = x\} f_1(x)\, dx \qquad \text{by (4.1.3)}$$

$$= \int_{-\infty}^{\infty} g(x) f_1(x)\, dx$$

in agreement with the previous intuitive argument.

4.2 EXAMPLES

We apply the general results of this section to some typical special cases.

▶ **Example 1.** A point is chosen with uniform density between 0 and 1. If the number R_1 selected is x, then a coin with probability x of heads is tossed independently n times. If R_2 is the resulting number of heads, find $p_2(k) = P\{R_2 = k\}$, $k = 0, 1, \dots, n$.

Here we have $f_1(x) = 1$, $0 \le x \le 1$; $f_1(x) = 0$ elsewhere. Also

$$P_x\{k\} = P\{R_2 = k \mid R_1 = x\} = \binom{n}{k} x^k (1 - x)^{n-k}$$

By (4.1.3),

$$P\{R_2 = k\} = \int_0^1 \binom{n}{k} x^k (1 - x)^{n-k}\, dx$$

This is an instance of the *beta function*, defined by

$$\beta(r, s) = \int_0^1 x^{r-1} (1 - x)^{s-1}\, dx, \qquad r, s > 0$$

It can be shown that the beta function can be expressed in terms of the gamma function [see (3.2.2)] by

$$\beta(r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r + s)} \tag{4.2.1}$$

(see Problem 1). Thus

$$p_2(k) = \binom{n}{k} \beta(k + 1, n - k + 1)$$

$$= \binom{n}{k} \frac{\Gamma(k + 1)\Gamma(n - k + 1)}{\Gamma(n + 2)}$$

$$= \binom{n}{k} \frac{k!\,(n - k)!}{(n + 1)!} = \frac{1}{n + 1}, \qquad k = 0, 1, \dots, n \quad ◀$$

▶ **Example 2.** A nonnegative number R_1 is chosen with the density $f_1(x) = xe^{-x}, x \geq 0; f_1(x) = 0, x < 0$. If $R_1 = x$, a number R_2 is chosen with uniform density between 0 and x. Find $P\{R_1 + R_2 \leq 2\}$.

Now we must have $0 \leq R_2 \leq R_1$; hence, if $0 \leq R_1 \leq 1$, then necessarily $R_1 + R_2 \leq 2$. If $1 < R_1 \leq 2$, then $R_1 + R_2 \leq 2$ provided that $R_2 \leq 2 - R_1$. If $R_1 > 2$, then $R_1 + R_2$ cannot be ≤ 2. By (4.1.2),

$$\begin{aligned} P\{R_1 + R_2 \leq 2\} & \\ &= \int_0^\infty xe^{-x} P\{R_1 + R_2 \leq 2 \mid R_1 = x\}\, dx \\ &= \int_0^1 xe^{-x}(1)\, dx + \int_1^2 xe^{-x} P\{R_2 \leq 2 - x \mid R_1 = x\} dx + \int_2^\infty xe^{-x}(0)\, dx \end{aligned}$$

Given $R_1 = x$, R_2 is uniformly distributed between 0 and x; thus

$$P\{R_2 \leq 2 - x \mid R_1 = x\} = \frac{2 - x}{x}, \qquad 1 \leq x \leq 2$$

(see Figure 4.2.1). Therefore

$$P\{R_1 + R_2 \leq 2\} = \int_0^1 xe^{-x}\, dx + \int_1^2 xe^{-x}\left(\frac{2 - x}{x}\right) dx = 1 - 2e^{-1} + e^{-2} \blacktriangleleft$$

▶ **Example 3.** Let R_1 be a discrete random variable, taking on the values $x_1, x_2, \ldots$ with probabilities $p(x_1), p(x_2), \ldots$. If $R_1 = x_i$, a random variable R_2 is observed, where R_2 has density f_i. What is $P\{(R_1, R_2) \in C\}$?

This is not quite the situation we considered in Section 4.1, since R_1 is discrete. However, the theorem of total probability should still be in force. R_1 takes the value x_i with probability $p(x_i)$; given that $R_1 = x_i$, the probability that $R_2 \in B$ is $P_{x_i}(B) = \int_B f_i(y)\, dy$. Thus we should have

$$P\{R_1 \in A, R_2 \in B\} = \sum_{x_i \in A} p(x_i) \int_B f_i(y)\, dy \tag{4.2.2}$$

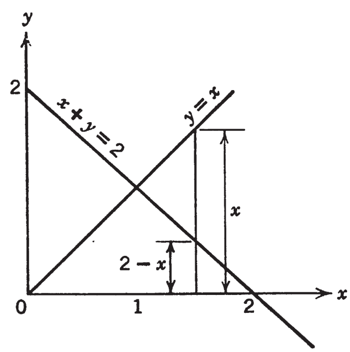

FIGURE 4.2.1 Conditional Probability Calculation.

and, more generally,

$$P\{(R_1, R_2) \in C\} = \sum_{x\,i} p(x_i) \int_{C_{x_i}} f_i(y)\, dy \tag{4.2.3}$$

In fact, if we take $\Omega = E^2$, $\mathscr{F} =$ Borel sets, $R_1(x, y) = x$, $R_2(x, y) = y$, it turns out that there is a unique probability measure on $\mathscr{F}$ satisfying (4.2.2) for all Borel subsets A, B of E^1; P is given explicitly by (4.2.3). ◀

PROBLEMS

1. Derive formula (4.2.1). HINT: in $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}\,dt$, let $t = x^2$. Then write $\Gamma(r)\Gamma(s)$ as a double integral and switch to polar coordinates.
2. In Example 2, what are the sets C and C_x in (4.1.2)? What is $P_x(C_x)$?
3. In Example 3, suppose that R_1 takes on positive integer values $1, 2, \ldots$ with probabilities $p_1, p_2, \ldots$ ($p_i \geq 0, \sum_{i=1}^\infty p_i = 1$). If $R_1 = n$, R_2 is selected according to the density $f_n(x) = ne^{-nx}$, $x \geq 0$; $f_n(x) = 0$, $x < 0$. Find the probability that $4 \leq R_1 + R_2 \leq 6$.
4. In Example 3 we specified $P_{x_i}(B)$ to be interpreted intuitively as the probability that $R_2 \in B$, given that $R_1 = x_i$. This, plus the specification of $p(x_i)$, $i = 1, 2, \ldots,$ determines the probability measure P. Use (4.2.2) to show that if $p(x_i) > 0$ then $P\{R_2 \in B \mid R_1 = x_i\} = P_{x_i}(B)$, thus justifying the intuition. In order words, the conditional probability as computed from the probability measure P coincides with the original specification.
5. A number R_1 is chosen with density $f_1(x) = 1/x^2$, $x \geq 1$; $f_1(x) = 0$, $x < 1$. If $R_1 = x$, let R_2 be uniformly distributed between 0 and x. Find the distribution and density functions of R_2.

4.3 CONDITIONAL DENSITY FUNCTIONS

We have seen that specification of the distribution or density function of a random variable R_1, together with $P_x(B)$ (for all real x and Borel subsets B of E^1), interpreted intuitively as the conditional probability that $R_2 \in B$, given $R_1 = x$, determines the probability of all events of the form $\{(R_1, R_2) \in C\}$. However, this has not resolved the difficulty of defining conditional probabilities given events of probability 0. If we are *given* random variables R_1 and R_2 with a particular joint distribution function, we can ask whether it is possible to define in a meaningful way the conditional probability $P\{R_2 \in B \mid R_1 = x\}$, even though the event $\{R_1 = x\}$ may have probability 0 for some, in fact perhaps for all, x. We now consider this question in the case in which R_1 and R_2 have a joint density f.

A reasonable approach to the conditional probability $P\{R_2 \in B \mid R_1 = x_0\}$ is to look at $P\{R_2 \in B \mid x_0 - h < R_1 < x_0 + h\}$ and let $h \to 0$. Now

$$P\{x_0 - h < R_1 < x_0 + h, R_2 \in B\} = \int_{x_0-h}^{x_0+h} \int_B f(x, y)\, dy\, dx$$

which for small h should look like $2h \int_B f(x_0, y)\, dy$. But $P\{x_0 - h < R_1 < x_0 + h\}$ looks like $2h\, f_1(x_0)$ for small h, where $f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$ is the density of R_1. Thus, as $h \to 0$, it appears that under appropriate conditions $P\{R_2 \in B \mid x - h < R_1 < x + h\}$ should approach $\int_B [f(x, y)/f_1(x)]\, dy$, so that we find conditional probabilities involving R_2, given $R_1 = x$, by integrating $f(x, y)/f_1(x)$ with respect to y.

We are led to define the *conditional density* of R_2 given $R_1 = x$ (or, for short, the conditional density of R_2 given R_1) as

$$h(y \mid x) = \frac{f(x, y)}{f_1(x)} \tag{4.3.1}$$

Since $\int_{-\infty}^{\infty} f(x, y)\, dy = f_1(x)$ (see Section 2.7), we have $\int_{-\infty}^{\infty} h(y \mid x)\, dy = 1$, so that $h(y \mid x)$, regarded as a function of y, is a legitimate density.

Notice that the conditional density is defined only when $f_1(x) > 0$. However, we may essentially ignore those (x, y) at which the conditional density is not defined. For let $S = \{(x, y): f_1(x) = 0\}$. We can show that $P\{(R_1, R_2) \in S\} = 0$.

$$P\{(R_1, R_2) \in S\} = \iint_S f(x, y)\, dx\, dy = \int_{\{x: f_1(x)=0\}} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx$$
$$= \int_{\{x: f_1(x)=0\}} f_1(x)\, dx = 0$$

We define the conditional probability that R_2 belongs to the Borel set B, given that $R_1 = x$, as

$$P_x(B) = P\{R_2 \in B \mid R_1 = x\} = \int_B h(y \mid x)\, dy \tag{4.3.2}$$

We can ask whether this is a sensible definition of conditional probability. We have set up our own ground rules to answer this question: "sensible" means that the theorem of total probability holds. Let us check that in fact (4.1.1) [and hence (4.1.2)] holds. We have

$$P\{R_1 \in A, R_2 \in B\} = \int_{x \in A} \int_{y \in B} f(x, y)\, dx\, dy$$
$$= \int_{x \in A} f_1(x) \left[\int_{y \in B} h(y \mid x)\, dy \right] dx = \int_A P_x(B) f_1(x)\, dx$$

which is (4.1.1).

We have seen that if (R_1, R_2) has density $f(x, y)$ and R_1 has density $f_1(x)$ we have a conditional density $h(y \mid x) = f(x, y)/f_1(x)$ for R_2, given $R_1 = x$. Let us reverse this process. Suppose that we observe a random variable R_1 with density $f_1(x)$; if $R_1 = x$, we observe a random variable R_2 with density $h(y \mid x)$. If we accept the continuous version of the theorem of total probability, we may calculate the joint distribution function of R_1 and R_2 using (4.1.1).

$$F(x_0, y_0) = P\{R_1 \le x_0, R_2 \le y_0\} = \int_{-\infty}^{x_0} P\{R_2 \le y_0 \mid R_1 = x\} f_1(x)\, dx$$

$$= \int_{-\infty}^{x_0} \left[\int_{-\infty}^{y_0} h(y \mid x)\, dy \right] f_1(x)\, dx = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} f_1(x) h(y \mid x)\, dy\, dx$$

Thus (R_1, R_2) has a density given by $f(x, y) = f_1(x)h(y \mid x)$, in agreement with (4.3.1).

To summarize: We may look at the formula $f(x, y) = f_1(x)h(y \mid x)$ in two ways.

1. If (R_1, R_2) has density $f(x, y)$, we have a natural notion of conditional probability.

$$P_x(B) = P\{R_2 \in B \mid R_1 = x\} = \int_B h(y \mid x)\, dy$$

2. If R_1 has density $f_1(x)$, and whenever $R_1 = x$ we select R_2 with density $h(y \mid x)$, then in the natural formulation of this problem (R_1, R_2) has density $f(x, y) = f_1(x)h(y \mid x)$.

In both cases "natural" indicates that (4.1.1), the continuous version of the theorem of total probability, is required to hold.

We may extend these results to higher dimensions. For example, if (R_1, R_2, R_3, R_4) has density $f(x_1, x_2, x_3, x_4)$, we define (say) the conditional density of (R_3, R_4) given (R_1, R_2), as

$$h(x_3, x_4 \mid x_1, x_2) = \frac{f(x_1, x_2, x_3, x_4)}{f_{12}(x_1, x_2)}$$

where

$$f_{12}(x_1, x_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_3\, dx_4$$

The conditional probability that (R_3, R_4) belongs to the two-dimensional Borel set B, given that $R_1 = x_1$, $R_2 = x_2$, is defined by

$$P_{x_1x_2}(B) = P\{(R_3, R_4) \in B \mid R_1 = x_1, R_2 = x_2\}$$

$$= \iint_B h(x_3, x_4 \mid x_1, x_2)\, dx_3\, dx_4$$

The appropriate version of the theorem of total probability is

$$P\{(R_1, R_2) \in A, (R_3, R_4) \in B\} = \iint_A P_{x_1x_2}(B) f_{12}(x_1, x_2)\, dx_1\, dx_2$$

If (R_1, R_2) has density $f_{12}(x_1, x_2)$, and having observed $R_1 = x_1$, $R_2 = x_2$, we select (R_3, R_4) with density $h(x_3, x_4 \mid x_1, x_2)$, then (R_1, R_2, R_3, R_4) must have density $f(x_1, x_2, x_3, x_4) = f_{12}(x_1, x_2)h(x_3, x_4 \mid x_1, x_2)$.

Let us do some examples.

▶ **Example 1.** We arrive at a bus stop at time $t = 0$. Two buses A and B are in operation. The arrival time R_1 of bus A is uniformly distributed between 0 and t_A minutes, and the arrival time R_2 of bus B is uniformly distributed between 0 and t_B minutes, with $t_A \le t_B$. The arrival times are independent. Find the probability that bus A will arrive first.

We are looking for the probability that $R_1 < R_2$. Since R_1 and R_2 are independent (and have a joint density), the conditional density of R_2 given R_1 is

$$\frac{f(x, y)}{f_1(x)} = f_2(y) = \frac{1}{t_B}, \qquad 0 \le y \le t_B$$

If bus A arrives at x, $0 \le x \le t_A$, it will be first provided that bus B arrives between x and t_B. This happens with probability $(t_B - x)/t_B$. Thus

$$P\{R_1 < R_2 \mid R_1 = x\} = 1 - \frac{x}{t_B}, \qquad 0 \le x \le t_A$$

By (4.1.2),

$$\begin{aligned} P\{R_1 < R_2\} &= \int_{-\infty}^{\infty} P\{R_1 < R_2 \mid R_1 = x\} f_1(x)\, dx \\ &= \int_0^{t_A} \left(1 - \frac{x}{t_B}\right)\frac{1}{t_A}\, dx = 1 - \frac{t_A}{2t_B} \end{aligned}$$

[Formally, taking the sample space as E^2, we have $C = \{R_1 < R_2\} = \{(x, y)\colon x < y\}$, $C_x = \{y\colon x < y\}$, $P_x(C_x) = P\{R_1 < R_2 \mid R_1 = x\} = 1 - x/t_B$, $0 \le x \le t_A$.]

Alternatively, we may simply use the joint density:

$$\begin{aligned} P\{R_1 < R_2\} &= \iint_{x<y} f(x, y)\, dx\, dy \\ &= \text{the shaded area in Figure 4.3.1, divided by the total area } t_At_B \\ &= 1 - \frac{t_A^2/2}{t_At_B} = 1 - \frac{t_A}{2t_B} \end{aligned}$$

as before. ◀

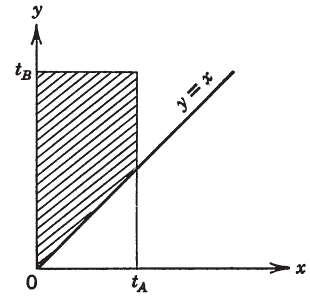

FIGURE 4.3.1 Bus Problem.

▶ **Example 2.** Let R_0 be a nonnegative random variable with density $f_0(\lambda) = e^{-\lambda}$, $\lambda \geq 0$. If $R_0 = \lambda$, we take n independent observations $R_1, R_2, \ldots, R_n$, each R_i having the exponential density $f_\lambda(y) = \lambda e^{-\lambda y}$, $y \geq 0$ ($= 0$ for $y < 0$). Find the conditional density of R_0 given $(R_1, R_2, \ldots, R_n)$.

Here we have specified $f_0(\lambda)$, the density of R_0, and the conditional density of $(R_1, R_2, \ldots, R_n)$ given R_0, namely,

$$h(x_1, x_2, \ldots, x_n \mid \lambda) = f_\lambda(x_1)f_\lambda(x_2)\cdots f_\lambda(x_n) \qquad \text{by the independence assumption}$$

$$= \lambda^n e^{-\lambda x}, \qquad x = \sum_{i=1}^{n} x_i$$

The joint density of $R_0, R_1, \ldots, R_n$ is therefore

$$f(\lambda, x_1, \ldots, x_n) = f_0(\lambda)h(x_1, \ldots, x_n \mid \lambda) = \lambda^n e^{-\lambda(1+x)}$$

The joint density of $R_1, \ldots, R_n$ is given by

$$g(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} f(\lambda, x_1, \ldots, x_n)\, d\lambda = \int_0^\infty \lambda^n e^{-\lambda(1+x)}\, d\lambda$$

$$= (\text{with } y = \lambda(1 + x)) \int_0^\infty \frac{y^n e^{-y}}{(1 + x)^{n+1}}\, dy = \frac{n!}{(1 + x)^{n+1}}$$

Thus the conditional density of R_0 given $(R_1, \ldots, R_n)$ is

$$h(\lambda \mid x_1, \ldots, x_n) = \frac{f(\lambda, x_1, \ldots, x_n)}{g(x_1, \ldots, x_n)} = \frac{1}{n!} \lambda^n e^{-\lambda(1+x)}(1 + x)^{n+1},$$

$$\lambda, x_1, \ldots, x_n \geq 0,\ x = x_1 + \cdots + x_n \blacktriangleleft$$

PROBLEMS

1. Let (R_1, R_2) have density $f(x, y) = e^{-y}$, $0 \leq x \leq y$, $f(x, y) = 0$ elsewhere. Find the conditional density of R_2 given R_1, and $P\{R_2 \leq y \mid R_1 = x\}$, the *conditional distribution function* of R_2 given $R_1 = x$.

2. Let (R_1, R_2) have density $f(x, y) = k\,|x|$, $-1 \le x \le 1$, $-1 \le y \le x$; $f(x, y) = 0$ elsewhere. Find k; also find the individual densities of R_1 and R_2, the conditional density of R_2 given R_1, and the conditional density of R_1 given R_2.

3. (a) If (R_1, R_2) is uniformly distributed over the set $C = \{(x, y): x^2 + y^2 \le 1\}$, show that, given $R_1 = x$, R_2 is uniformly distributed between $-(1 - x^2)^{1/2}$ and $+(1 - x^2)^{1/2}$.

(b) Let (R_1, R_2) be uniformly distributed over the arbitrary two-dimensional Borel set C [i.e., $P(B) =$ (area of $B \cap C$)/area of C ($=$ area B/area C if $B \subset C$)].

Show that given $R_1 = x$, R_2 is uniformly distributed on $C_x = \{y: (x, y) \in C\}$. In other words, $h(y \mid x)$ is constant for $y \in C_x$, and 0 for $y \notin C_x$.

4. In Problem 1, let $R_3 = R_2 - R_1$. Find the conditional density of R_3 given $R_1 = x$. Also find $P\{1 \le R_3 \le 2 \mid R_1 = x\}$.

5. Suppose that (R_1, R_2) has density f and $R_3 = g(R_1, R_2)$. You are asked to compute the conditional distribution function of R_3, given $R_1 = x$; that is, $P\{R_3 \le z \mid R_1 = x\}$. How would you go about it?

4.4 CONDITIONAL EXPECTATION

In the preceding sections we considered situations in which two successive observations are made, the second observation depending on the result of the first. The essential ingredient in such problems is the quantity $P_x(B)$, defined for real x and Borel sets B, to be interpreted as the conditional probability that the second observation will fall into B, given that the first observation takes the value x: for short, $P\{R_2 \in B \mid R_1 = x\}$. In particular, we may define the *conditional distribution function* of R_2 given $R_1 = x$, as $F_2(y \mid x) = P\{R_2 \le y \mid R_1 = x\}$.

If R_1 and R_2 have a joint density, this can be computed from the conditional density of R_2 given R_1: $F_2(y_0 \mid x) = \int_{-\infty}^{y_0} h(y \mid x)\,dy$.

In any case, for each real x we have a probability measure P_x defined on the Borel subsets of E^1. Now if $R_1 = x$ and we observe R_2, there should be an average value associated with R_2, that is, a conditional expectation of R_2 given that $R_1 = x$. How should this be computed? Let us try to set up an appropriate model. We are observing a single random variable R_2, so let $\Omega = E^1$, $\mathscr{F} =$ Borel sets, $R_2(y) = y$. We are not concerned with the probability that $R_2 \in B$, but instead with the probability that $R_2 \in B$, *given that* $R_1 = x$. In other words, the appropriate probability measure is P_x. The expectation of R_2, computed with respect to P_x, is called the *conditional expectation of* R_2 *given that* $R_1 = x$ (or, for short, the conditional expectation of R_2 given R_1), written $E(R_2 \mid R_1 = x)$.

Note that if g is a (piecewise continuous) function from E^1 to E^1, then $g(R_2)$ is also a random variable (see Section 2.7), so that we may also talk about the conditional expectation of $g(R_2)$ given $R_1 = x$, written $E[g(R_2) \mid R_1 = x]$. In particular, if there is a conditional density of R_2 given $R_1 = x$, then, by Theorem 2 of Section 3.1,

$$E[g(R_2) \mid R_1 = x] = \int_{-\infty}^{\infty} g(y)h(y \mid x)\,dy \tag{4.4.1}$$

if $g \geq 0$ or if the integral is absolutely convergent.

There is an immediate extension to n dimensions. For example, if there is a conditional density of (R_4, R_5) given (R_1, R_2, R_3), then

$$E[g(R_4, R_5) \mid R_1 = x_1, R_2 = x_2, R_3 = x_3]$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_4, x_5)h(x_4, x_5 \mid x_1, x_2, x_3)\,dx_4\,dx_5$$

Note also that conditional probability can be obtained from conditional expectation. If in (4.4.1) we take $g(y) = I_B(y) = 1$ if $y \in B$, and $= 0$ if $y \notin B$, then

$$E[g(R_2) \mid R_1 = x] = E[I_B(R_2) \mid R_1 = x] = \int_{-\infty}^{\infty} I_B(y)h(y \mid x)\,dy$$
$$= \int_B h(y \mid x)\,dy = P\{R_2 \in B \mid R_1 = x\}$$

We have seen previously that $P\{R_2 \in B\} = E[I_{\{R_2 \in B\}}]$. We now have a similar result under the condition that $R_1 = x$. [Notice that $I_B(R_2) = I_{\{R_2 \in B\}}$; for $I_B(R_2(\omega)) = 1$ iff $R_2(\omega) \in B$, that is, iff $I_{\{R_2 \in B\}}(\omega) = 1$.]

Let us consider again the examples of Section 4.2.

▶ **Example 1.** R_1 is uniformly distributed between 0 and 1; if $R_1 = x$, R_2 is the number of heads in n tosses of a coin with probability x of heads.

Given that $R_1 = x$, R_2 has a binomial distribution with parameters n and x: $P\{R_2 = k \mid R_1 = x\} = \binom{n}{k}x^k(1-x)^{n-k}$. It follows that $E(R_2 \mid R_1 = x)$ is the average number of successes in n Bernoulli trials, with probability x of success on a particular trial, namely, nx. ◀

▶ **Example 2.** R_1 has density $f_1(x) = xe^{-x}$, $x \geq 0$, $f_1(x) = 0$, $x < 0$. The conditional density of R_2 given $R_1 = x$ is uniform over $[0, x]$. It follows that, for $x > 0$,

$$E(R_2 \mid R_1 = x) = \int_{-\infty}^{\infty} yh(y \mid x)\,dy = \int_0^x y\frac{1}{x}\,dy = \tfrac{1}{2}x$$

Similarly,

$$E[e^{R_2} \mid R_1 = x] = \int_{-\infty}^{\infty} e^y h(y \mid x)\,dy = \int_0^x e^y \frac{1}{x}\,dy = \frac{e^x - 1}{x} \quad ◀$$

► **Example 3.** R_1 is discrete, with $p(x_i) = P\{R_1 = x_i\}$, $i = 1, 2, \ldots$. Given $R_1 = x_i$, R_2 has density f_i; that is,

$$P\{R_2 \in B \mid R_1 = x_i\} = \int_B f_i(y)\, dy$$

Thus

$$E[g(R_2) \mid R_1 = x_i] = \int_{-\infty}^{\infty} g(y) f_i(y)\, dy \quad ◄$$

Now let us consider a slightly different case.

► **Example 4.** Let R_1 and R_2 be discrete random variables. If $R_1 = x$, then R_2 will take the value y with probability

$$p(y \mid x) = P\{R_2 = y \mid R_1 = x\} = \frac{p_{12}(x, y)}{p_1(x)}$$

where

$$p_{12}(x, y) = P\{R_1 = x, R_2 = y\},\ p_1(x) = P\{R_1 = x\}$$

$p(y \mid x)$, which is defined provided that $p_1(x) > 0$, will be called the *conditional probability function* of R_2 given $R_1 = x$ (or the conditional probability function of R_2 given R_1, for short). We may find the probability that $R_2 \in B$ given $R_1 = x$ by summing the conditional probability function.

$$P_x(B) = P\{R_2 \in B \mid R_1 = x\} = \frac{P\{R_1 = x, R_2 \in B\}}{P\{R_1 = x\}} = \frac{\sum_{y \in B} p_{12}(x, y)}{p_1(x)}$$
$$= \sum_{y \in B} p(y \mid x)$$

Thus, given that $R_1 = x$, the probabilities of events involving R_2 are found from the probability function $p(y \mid x)$, y real. Therefore the conditional expectation of $g(R_2)$ given $R_1 = x$ is

$$E[g(R_2) \mid R_1 = x] = \sum_y g(y) p(y \mid x) \tag{4.4.2}$$

In particular,

$$E(R_2 \mid R_1 = x) = \sum_y y p(y \mid x) \quad ◄$$

There is a feature common to all these examples. In each case the expectation of R_2 (or of a function of R_2) can be expressed as a weighted average of conditional expectations. Let us look at Example 4 first. With probability $p_1(x)$, R_1 takes the value x; if $R_1 = x$, the average value of R_2 is $E(R_2 \mid R_1 = x)$. By analogy with the theorem of total probability, it is reasonable to expect that

$$E(R_2) = \sum_x p_1(x) E(R_2 \mid R_1 = x)$$

To justify this, write

$$E(R_2) = \sum_y y p_2(y) = \sum_y y P\{R_2 = y\} = \sum_y y \sum_x P\{R_1 = x, R_2 = y\} \qquad \text{by (2.7.2)}$$

$$= \sum_{x,y} y P\{R_1 = x\} P\{R_2 = y \mid R_1 = x\} = \sum_x p_1(x) [\sum_y y p(y \mid x)]$$

This is the desired result.

In Example 1 the probability that R_1 will lie in an interval about x is $f_1(x)\,dx = dx$; given that $R_1 = x$, the average value of R_2 is $E(R_2 \mid R_1 = x) = nx$. We expect that

$$E(R_2) = \int_{-\infty}^{\infty} f_1(x) E(R_2 \mid R_1 = x)\,dx$$

To verify this, notice that we calculated in Section 4.2 that

$$P\{R_2 = k\} = \frac{1}{n+1}, \qquad k = 0, 1, \ldots, n$$

Thus

$$E(R_2) = \sum_{k=0}^{n} k P\{R_2 = k\} = \frac{1}{n+1}(1 + 2 + \cdots + n) = \frac{1}{n+1}\frac{(n+1)n}{2} = \frac{n}{2}$$

But

$$\int_{-\infty}^{\infty} f_1(x) E(R_2 \mid R_1 = x)\,dx = \int_0^1 nx\,dx = \frac{n}{2}$$

In Example 2, the joint density of R_1 and R_2 is

$$f(x, y) = f_1(x) h(y \mid x) = \frac{xe^{-x}}{x} = e^{-x}, \qquad x \geq 0,\ 0 \leq y \leq x$$

Now

$$E(R_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x, y)\,dx\,dy$$

[*Notice that we need not compute $f_2(y)$ explicitly;* instead we simply regard R_2 as a function of R_1 and R_2; that is, we set $g(R_1, R_2) = R_2$ and compute

$$E[g(R_1, R_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f(x, y)\,dx\,dy]$$

Thus

$$E(R_2) = \int_0^{\infty} e^{-x}\left[\int_0^x y\,dy\right] dx = \int_0^{\infty} \tfrac{1}{2}x^2 e^{-x}\,dx = \tfrac{1}{2}\Gamma(3) = 1$$

But

$$\int_{-\infty}^{\infty} f_1(x) E(R_2 \mid R_1 = x)\,dx = \int_0^{\infty} xe^{-x}(\tfrac{1}{2}x)\,dx = 1$$

In Example 3 we have [see (4.2.2)]

$$P\{R_2 \in B\} = \sum_i p(x_i) \int_B f_i(y)\, dy = \int_B \Big[\sum_i p(x_i) f_i(y)\Big]\, dy$$

so that R_2 has density

$$f_2(y) = \sum_i p(x_i) f_i(y) \tag{4.4.3}$$

Thus

$$E(R_2) = \int_{-\infty}^{\infty} y f_2(y)\, dy = \sum_i p(x_i) \int_{-\infty}^{\infty} y f_i(y)\, dy$$

and consequently

$$E(R_2) = \sum_i p(x_i) E(R_2 \mid R_1 = x_i)$$

as expected.

Results of the form

$$E(R_2) = \sum_i p(x_i) E(R_2 \mid R_1 = x_i) \tag{4.4.4}$$

or

$$E(R_2) = \int_{-\infty}^{\infty} f_1(x) E(R_2 \mid R_1 = x)\, dx \tag{4.4.5}$$

are called versions of the *theorem of total expectation.*

In the situations we are considering, conditional expectations are derived ultimately from a given set of probabilities $P_x(B) = P\{R_2 \in B \mid R_1 = x\}$. In such cases it turns out that if $E(R_2)$ exists, (4.4.4) will hold if R_1 is discrete, and (4.4.5) will hold if R_1 is absolutely continuous.

Notice that $E(R_2 \mid R_1 = x)$ will in general depend on x and hence may be written as $g(x)$; $\int_{-\infty}^{\infty} g(x) f_1(x)\, dx$ in (4.4.5) [or $\sum_x g(x)p(x)$ in (4.4.4)] is then the expectation of $g(R_1)$. Thus (4.4.4) and (4.4.5) may be rephrased as follows.

The expectation of the conditional expectation of R_2 given R_1 is the (over-all) expectation of R_2.

▶ **Example 5.** Let R be a random variable with the distribution function shown in Figure 4.4.1. Find $E(R^3)$.

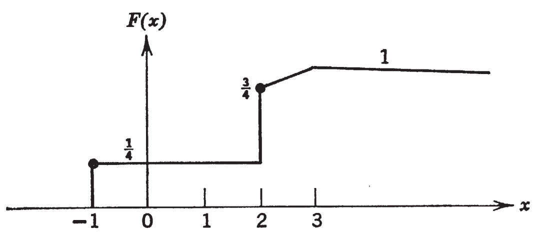

FIGURE 4.4.1

If R were discrete we would compute

$$E(R^3) = \sum_x x^3 p_R(x)$$

and if R were absolutely continuous we would compute

$$E(R^3) = \int_{-\infty}^{\infty} x^3 f_R(x)\, dx$$

In this case, however, R falls into neither category. We are going to show how to use the theorem of total expectation to compute $E(R^3)$.

We have $P\{R = -1\} = 1/4$, $P\{R = 2\} = 3/4 - 1/4 = 1/2$, $P\{R = x\} = 0$ for other values of x. Let F_1 be a step function that is 0 for $x < -1$ and has a jump of 1/4 at $x = -1$ and a jump of 1/2 at $x = 2$. Subtract F_1 from F to obtain a continuous function F_2 that can be represented as an integral of a nonnegative function f_2. F_1 is called the "discrete part" of F, and F_2 the "absolutely continuous part" (see Figure 4.4.2). F_1 and F_2 are monotone, right-continuous functions, and they approach zero as $x \to -\infty$. However, they approach limits that are less than 1 as $x \to \infty$, so that they cannot be regarded as distribution functions of random variables. However, $(4/3)F_1$ and $4F_2$ are legitimate distribution functions.

We shall show that

$$E(R^3) = \sum_x x^3 p_R(x) + \int_{-\infty}^{\infty} x^3 f_2(x)\, dx$$

Consider the following random experiment. With probability 3/4 $(= F_1(\infty) = \sum_x p_R(x)$, where $p_R(x) = P\{R = x\})$, pick a number in accordance

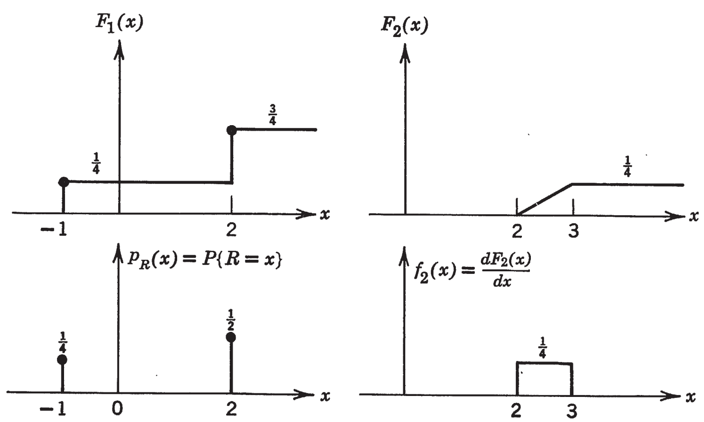

FIGURE 4.4.2 Discrete and Absolutely Continuous Parts of a Distribution Function.

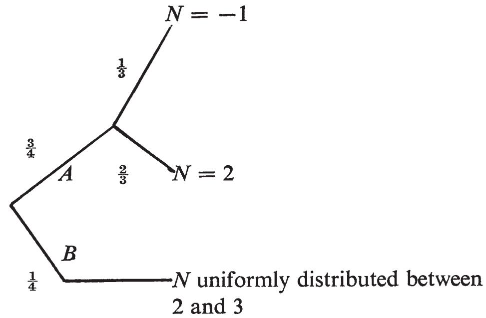

FIGURE 4.4.3 Tree Diagram for Example 5.

with $(4/3)F_1$; that is, pick -1 with probability 1/3 and 2 with probability 2/3. With probability 1/4 [$= F_2(\infty)$], pick a number in accordance with F_2, that is, one uniformly distributed between 2 and 3 (see Figure 4.4.3).

If N is the resulting number, then, by the theorem of total probability,

$$P\{N \le x\} = P(A)P\{N \le x \mid A\} + P(B)P\{N \le x \mid B\}$$

where A and B correspond to the two possible results at the first stage of the experiment. Thus

$$F_N(x) = \tfrac{3}{4}(\tfrac{4}{3}F_1(x)) + \tfrac{1}{4}(4F_2(x)) = F_1(x) + F_2(x) = F(x)$$

Therefore F_N is the original distribution function F.

Since N and R have the same distribution function, we expect that $E(N^3) = E(R^3)$. Now we may compute $E(N^3)$ by the theorem of total expectation.

$$\begin{aligned} E(N^3) &= P(A)E(N^3 \mid A) + P(B)E(N^3 \mid B) \\ &= \tfrac{3}{4}[(-1)^3\tfrac{1}{3} + 2^3\,\tfrac{2}{3}] + \tfrac{1}{4}\int_2^3 x^3\,dx = \tfrac{15}{4} + \tfrac{65}{16} = \tfrac{125}{16} \end{aligned}$$

Notice that this may be expressed as

$$(-1)^3\,\tfrac{1}{4} + 2^3\,\tfrac{1}{2} + \int_2^3 x^3\,\tfrac{1}{4}\,dx$$

that is,

$$E(R^3) = \sum_x x^3 p_R(x) + \int_{-\infty}^{\infty} x^3 f_2(x)\,dx$$

More generally, the expectation of a function of R may be computed by

$$E[g(R)] = \sum_x g(x)p_R(x) + \int_{-\infty}^{\infty} g(x)f_2(x)\,dx \qquad (4.4.6)\dagger$$

if $g \geq 0$ or if both the series and the integral are absolutely convergent. ◀

▶ **Example 6.** Let R be a random variable on a given probability space, and A an event with $P(A) > 0$. Formulate the proper definition of the conditional expectation of R, given that A has occurred.

This actually is not a new concept. If I_A is the indicator of A, we are looking for the expectation of R, given that $I_A = 1$. Let the experiment be performed independently n times, n very large, and let R_i be the value of R obtained on trial i, $i = 1, 2, \ldots, n$. Renumber the trials so that A occurs on the first k trials, and A^c on the last $n - k$ [k will be approximately $nP(A)$]. The average value of R, considering only those trials on which A occurs, is

$$\frac{R_1 + \cdots + R_k}{k} = \left(\frac{1}{n}\sum_{j=1}^{n} R_j I_j\right)\frac{n}{k}$$

where $I_j = 1$ if A occurs on trial j; $I_j = 0$ if A does not occur on trial j. In other words, I_j is simply the jth observation of I_A. It appears that $1/n \sum_{j=1}^{n} R_j I_j$ approximates the expectation of RI_A; since k/n approximates $P(A)$, we are led to define the *conditional expectation of R given A* as

$$E(R \mid A) = \frac{E(RI_A)}{P(A)} \qquad \text{if } P(A) > 0 \qquad (4.4.7)$$

Let us check that (4.4.7) agrees with previous results when R is discrete. By (4.4.2),

$$E(R \mid I_A = 1) = \sum_y yP\{R = y \mid I_A = 1\} = \sum_{y \neq 0} yP\{R = y \mid I_A = 1\}$$

But if $y \neq 0$,

$$P\{R = y \mid I_A = 1\} = \frac{P\{R = y, I_A = 1\}}{P\{I_A = 1\}} = \frac{P\{RI_A = y\}}{P(A)}$$

Thus

$$E(R \mid I_A = 1) = \frac{1}{P(A)} \sum_{y \neq 0} yP\{RI_A = y\} = \frac{E(RI_A)}{P(A)}$$

† The reader may recognize this as the Riemann-Stieltjes integral $\int_{-\infty}^{\infty} g(x)\,dF(x)$. Alternatively, if one differentiates F formally to obtain $f = f_2$ plus "impulses" or "delta functions" at -1 and 2 of strength 1/4 and 1/2, respectively, and then evaluates $\int_{-\infty}^{\infty} g(x)f(x)\,dx$, (4.4.6) is obtained.

Let us look at another special case. For any random variable R and event A with $P(A) > 0$, we may define the *conditional distribution function of R given A* in a natural way, namely,

$$F_R(x \mid A) = P\{R \le x \mid A\} = \frac{P(A \cap \{R \le x\})}{P(A)} \tag{4.4.8}$$

Now assume that R has density f and A is of the form $\{R \in B\}$ for some Borel set B. Then

$$P(A \cap \{R \le x_0\}) = P\{R \in B, R \le x_0\} = \int_{\substack{x \in B \\ x \le x0}} f(x)\, dx$$

$$= \int_{x \le x_0} f(x) I_B(x)\, dx$$

Thus (4.4.8) becomes

$$F_R(x_0 \mid A) = \int_{-\infty}^{x_0} \frac{f(x)}{P(A)} I_B(x)\, dx$$

In other words, there is a *conditional density of R given A*, namely,

$$\begin{aligned} f_R(x \mid A) = \frac{f(x)}{P(A)} I_B(x) &= \frac{f(x)}{P(A)} \qquad \text{if } x \in B \\ &= 0 \qquad \text{if } x \notin B \end{aligned} \tag{4.4.9}$$

We may then compute the conditional expectation of R given A.

$$E(R \mid A) = \int_{-\infty}^{\infty} x f_R(x \mid A)\, dx = \int_{-\infty}^{\infty} \frac{x I_B(x)}{P(A)} f(x)\, dx$$

$$= \frac{E(R I_B(R))}{P(A)} \qquad \text{by Theorem 2 of Section 3.1}$$

But

$$\begin{aligned} I_B(R) &= I_{\{R \in B\}} \qquad \text{by the discussion preceding Example 1} \\ &= I_A \end{aligned}$$

Thus

$$E(R \mid A) = \frac{E(R I_A)}{P(A)}$$

in agreement with (4.4.7).

REMARK. (4.4.8) and (4.4.9) extend to n dimensions. The conditional distribution function of $(R_1, \ldots, R_n)$ given A is $F_{12\cdots n}(x_1, \ldots, x_n \mid A) = P\{R_1 \le x_1, \ldots, R_n \le x_n \mid A\}$. If $(R_1, \ldots, R_n)$ has density f and $A = \{(R_1, \ldots, R_n) \in B\}$, there is a conditional density of

$(R_1, \ldots, R_n)$ given A.

$$f_R(x_1, \ldots, x_n \mid A) = \frac{f(x_1, \ldots, x_n)}{P(A)} I_B(x_1, \ldots, x_n)$$

The argument is essentially the same as above. ◀

PROBLEMS

1. Let (R_1, R_2) have density $f(x, y) = 8xy, 0 \le y \le x \le 1; f(x, y) = 0$ elsewhere.
   (a) Find the conditional expectation of R_2 given $R_1 = x$, and the conditional expectation of R_1 given $R_2 = y$.
   (b) Find the conditional expectation of $R_2{}^4$ given $R_1 = x$.
   (c) Find the conditional expectation of R_2 given $A = \{R_1 \le 1/2\}$.
2. In Example 2 of Section 4.3, find the conditional expectation of $R_0{}^{-n}$, given $R_1 = x_1, \ldots, R_n = x_n$.
3. Let (R_1, R_2) be uniformly distributed over the parallelogram with vertices $(0, 0)$, $(2, 0)$, $(3, 1)$, $(1, 1)$. Find $E(R_2 \mid R_1 = x)$.
4. If a single die is tossed independently n times, find the average number of 2's, given that the number of 1's is k.
5. Let R_1 and R_2 be independent random variables, each uniformly distributed between 0 and 2.
   (a) Find the conditional probability that $R_1 \ge 1$, given that $R_1 + R_2 \le 3$.
   (b) Find the conditional expectation of R_1, given that $R_1 + R_2 \le 3$.
6. Let $B_1, B_2, \ldots$ be mutually exclusive, exhaustive events, with $P(B_n) > 0$, $n = 1, 2, \ldots$, and let R be a random variable. Establish the following version of the theorem of total expectation:

$$E(R) = \sum_{n=1}^{\infty} P(B_n)E(R \mid B_n)$$

   [if $E(R)$ exists].
7. Of the 100 people in a certain village, 50 always tell the truth, 30 always lie, and 20 always refuse to answer. A single unbiased die is tossed. If the result is 1, 2, 3, or 4, a sample of size 30 is taken *with replacement*. If the result is 5 or 6, a sample of size 30 is taken *without replacement*. A random variable R is defined as follows:
   $R = 1$ if the resulting sample contains 10 people of each category.
   $R = 2$ if the sample is taken with replacement and contains 12 liars.
   $R = 3$ otherwise.
   Find $E(R)$.

8. Let R_1 and R_2 be independent random variables, each uniformly distributed between 0 and 1. Define

$$R_3 = g(R_1, R_2) = R_1 \quad \text{if } R_1^2 + R_2^2 \le 1$$
$$R_3 = 2 \quad \text{if } R_1^2 + R_2^2 > 1$$

(a) Find $F_3(z)$ and compute $E(R_3)$ from this.
(b) Compute $E(R_3)$ from $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f_{12}(x, y)\,dx\,dy$.
(c) Compute $E(R_3 \mid R_1^2 + R_2^2 \le 1)$ and $E(R_3 \mid R_1^2 + R_2^2 > 1)$; then find $E(R_3)$ by using the theorem of total expectation.

9. The density for the time T required for the failure of a light bulb is $f(x) = \lambda e^{-\lambda x}$, $x \ge 0$. Find the conditional density function of $T - t_0$, given that $T > t_0$, and interpret the result intuitively.

10. Let R_1 and R_2 be independent random variables, each uniformly distributed between 0 and 1. Find the conditional expectation of $(R_1 + R_2)^2$ given $R_1 - R_2$.

11. Let R_1 and R_2 be independent random variables, each with density $f(x) = (1/2)e^{-x}$, $x \ge 0$; $f(x) = 1/2$, $-1 \le x \le 0$; $f(x) = 0$, $x < -1$. Let $R_3 = R_1^2 + R_2^2$. Find $E(R_3 \mid R_1 = x)$.

12. Let R_1 be a discrete random variable; if $R_1 = x$, let R_2 have a conditional density $h(y \mid x)$. Define the conditional probability that $R_1 = x$ given that $R_2 = y$ as

$$P\{R_1 = x \mid R_2 = y\} = \frac{P\{R_1 = x\}h(y \mid x)}{\sum_{x'} P\{R_1 = x'\}h(y \mid x')}$$

(cf. Bayes' Theorem).
(a) Interpret this definition intuitively by considering $P\{R_1 = x \mid y < R_2 < y + dy\}$.
(b) Show that the definition is natural in the sense that the appropriate version of the theorem of total probability is satisfied:

$$P\{R_1 \in A, R_2 \in B\} = \int_B f_2(y)P\{R_1 \in A \mid R_2 = y\}\,dy$$

where

$$P\{R_1 \in A \mid R_2 = y\} = \sum_{x \in A} P\{R_1 = x \mid R_2 = y\}$$
$$f_2(y) = \sum_{x} P\{R_1 = x\}h(y \mid x)$$

[see (4.4.3)].

13. If R_1 is absolutely continuous and R_2 discrete, and $p(y \mid x) = P\{R_2 = y \mid R_1 = x\}$ is specified, show that there is a conditional density of R_1 given R_2, namely,

$$h(x \mid y) = \frac{f_1(x)p(y \mid x)}{p_2(y)}$$

where

$$p_2(y) = P\{R_2 = y\} = \int_{-\infty}^{\infty} f_1(x)p(y \mid x)\,dx$$

14. Let R be uniformly distributed between 0 and 1. If $R = \lambda$, a coin with probability of heads λ is tossed independently n times. If $R_1, \dots, R_n$ are the results of the tosses ($R_i = 1$ for a head, $R_i = 0$ for a tail), find the conditional density of R given $(R_1, \dots, R_n)$, and the conditional expectation of R given $(R_1, \dots, R_n)$.

15. (Hypothesis testing) Consider the following experiment. Throw a coin with probability p of heads. If the coin comes up heads, observe a random variable R with density $f_0(x)$; if the coin comes up tails, let R have density $f_1(x)$. Suppose that we are not told the result of the coin toss, but only the value of R, and we have to guess whether or not the coin came up heads. We do this by means of a *decision scheme*, which is simply a Borel set S of real numbers with the interpretation that if $R = x$ and $x \in S$, we decide for tails, that is, f_1, and if $x \notin S$ we decide for heads, that is, f_0.

(a) Find the over-all probability of error in terms of p, f_0, f_1, and S. [There are two types of errors: if the actual density is f_0 and we decide for f_1 (type 1 error), and if the actual density is f_1 and we decide for f_0 (type 2 error).]

(b) For a given p, f_0, f_1, find the S that makes the over-all probability of error a minimum. Apply the results to the case in which f_i is the normal density with mean m_i and variance σ^2, $i = 0, 1$.

REMARK. A physical model for part (b) is the following. The input R to a radar receiver is of the form $\theta + N$, where θ (the signal) and N (the noise) are independent random variables, with $P\{\theta = m_0\} = p$, $P\{\theta = m_1\} = 1 - p$, and N normally distributed with mean 0 and variance σ^2. If $\theta = m_i$ ($i = 0$ corresponds to a head in the above discussion, and $i = 1$ to a tail), then R is normal with mean m_i and variance σ^2; thus f_i is the conditional density of R given $\theta = m_i$. We are trying to determine the actual value of the signal with as low a probability of error as possible.

16. Let R be the number of successes in n Bernoulli trials, with probability p of success on a given trial. Find the conditional expectation of R, given that $R \geq 2$.

17. Let R_1 be uniformly distributed between 0 and 10, and define R_2 by

$$\begin{aligned} R_2 &= R_1^2 \qquad \text{if } 0 \leq R_1 \leq 6 \\ &= 3 \qquad \text{if } 6 < R_1 \leq 10 \end{aligned}$$

Find the conditional expectation of R_2 given that $2 \leq R_2 \leq 4$.

18. Consider the following two-stage random experiment.

(i) A circle of radius R and center at $(0, 0)$ is selected, where R has density $f_R(z) = e^{-z}$, $z \geq 0$; $f_R(z) = 0$, $z < 0$.

(ii) A point (R_1, R_2) is chosen, where (R_1, R_2) is uniformly distributed inside the circle selected in step (i).

(a) If $D = (R_1^2 + R_2^2)^{1/2}$ is the distance of the resulting point from the origin, find $E(D)$.

(b) Find the conditional density of R given $R_1 = x$, $R_2 = y$. (Leave the answer in the form of an integral.)

19. (An estimation problem) The input R to a radar receiver is of the form $\theta + N$, where θ (the signal) and N (the noise) are independent random variables with finite mean and variance. The value of R is observed, and then an estimate of θ is made, say, $\theta^* = d(R)$, where d is a function from the reals to the reals. We wish to choose the estimate so that $E[(\theta^* - \theta)^2]$ is as small as possible.

(a) Show that $d(x)$ is the conditional expectation $E(\theta \mid R = x)$. (Assume that R is either absolutely continuous or discrete.)

(b) Let $\theta = \pm 1$ with equal probability, and let N be uniformly distributed between -2 and $+2$. Find $d(x)$ and the minimum value of $E[(\theta^* - \theta)^2]$.

20. A number θ is chosen at random with density $f_\theta(x) = e^{-x}$, $x \geq 0$; $f_\theta(x) = 0$, $x < 0$. If θ takes the value λ, a random variable R is observed, where R has the Poisson distribution with parameter λ. For example, R might be the number of radioactive particles (or particles with some other distinguishing characteristic) passing through a counting device in a given time interval, where the average number of such particles is selected randomly. The value of R is observed and an estimate of θ is made, say $\theta^* = d(R)$. The argument of Problem 19, which applies in any situation when one makes an estimate $\theta^* = d(R)$ of a parameter θ, and when the distribution function of R depends on θ, shows that the estimate that minimizes $E[(\theta^* - \theta)^2]$ is $d(x) = E(\theta \mid R = x)$. Find $d(x)$ in this case.

REMARK. Problems 15, 19, and 20 illustrate some techniques of statistics. This subject will be taken up systematically in Chapter 8.

4.5 APPENDIX: THE GENERAL CONCEPT OF CONDITIONAL EXPECTATION

By shifting our viewpoint slightly, we may regard a conditional expectation as a random variable defined on the given probability space. For example, suppose that $E(R_2 \mid R_1 = x) = x^2$. We may then say that, having observed R_1, the average value of R_2 is R_1^2. We adopt the notation $E(R_2 \mid R_1) = R_1^2$. In general, if $E(R_2 \mid R_1 = x) = g(x)$, we define $E(R_2 \mid R_1) = g(R_1)$. Then $E(R_2 \mid R_1)$ is a function defined on Ω; its value at the point ω is $g(R_1(\omega))$.

Let us see what happens to the theorem of total expectation in this notation. If, for example,

$$E(R_2) = \int_{-\infty}^{\infty} f_1(x) E(R_2 \mid R_1 = x)\, dx = \int_{-\infty}^{\infty} f_1(x) g(x)\, dx$$

then $E(R_2) = E[g(R_1)]$; in other words,

$$E(R_2) = E[E(R_2 \mid R_1)] \tag{4.5.1}$$

The expectation of the conditional expectation of R_2 given R_1 is the expectation of R_2.

Let us develop this a bit further. Let A be a Borel subset of E^1. Then, assuming that (4.5.1) holds for the random variable $R_2 I_{\{R_1 \in A\}}$, we have

$$E(R_2 I_{\{R_1 \in A\}}) = E[E(R_2 I_{\{R_1 \in A\}} \mid R_1)]$$

But having observed R_1, $R_2 I_{\{R_1 \in A\}}$ will be R_2 if $R_1 \in A$, and 0 otherwise; thus we expect intuitively that

$$E(R_2 I_{\{R_1 \in A\}} \mid R_1) = I_{\{R_1 \in A\}} E(R_2 \mid R_1)$$

It appears reasonable to expect, then, that

$$E(R_2 I_{\{R_1 \in A\}}) = E[I_{\{R_1 \in A\}} E(R_2 \mid R_1)] \qquad \text{for all Borel subsets } A \text{ of } E^1 \tag{4.5.2}$$

It turns out that if R_1 is an arbitrary random variable and R_2 a random variable whose expectation exists, there is a random variable R, of the form $g(R_1)$ for some Borel measurable function g, such that

$$E(R_2 I_{\{R_1 \in A\}}) = E[I_{\{R_1 \in A\}} R] \qquad \text{for all Borel subsets } A \text{ of } E^1$$

We set $R = E(R_2 \mid R_1)$. Furthermore, R is essentially unique: if $R' = g'(R_1)$ for some Borel measurable function g', and R' also satisfies (4.5.2), then $R = R'$ except perhaps on a set of probability 0.

In the cases considered in this chapter, the conditional expectations all satisfy (4.5.2) (which is just a restatement of the theorem of total expectation), and thus the examples of the chapter are consistent with the general notion of conditional expectation.