2.5 Jointly Distributed Random Variables

2.5.1 Joint Distribution Functions

Thus far, we have concerned ourselves with the probability distribution of a single random variable. However, we are often interested in probability statements concerning two or more random variables. To deal with such probabilities, we define, for any two random variables X and Y, the *joint cumulative probability distribution function* of X and Y by

$$F(a, b) = P\{X \leq a, Y \leq b\}, \qquad -\infty < a, b < \infty$$

The distribution of X can be obtained from the joint distribution of X and Y as follows:

$$\begin{aligned} F_X(a) &= P\{X \leq a\} \\ &= P\{X \leq a, Y < \infty\} \\ &= F(a, \infty) \end{aligned}$$

Similarly, the cumulative distribution function of Y is given by

$$F_Y(b) = P\{Y \leq b\} = F(\infty, b)$$

In the case where X and Y are both discrete random variables, it is convenient to define the *joint probability mass function* of X and Y by

$$p(x, y) = P\{X = x, Y = y\}$$

The probability mass function of X may be obtained from $p(x, y)$ by

$$p_X(x) = \sum_{y:p(x,y)>0} p(x, y)$$

Similarly,

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x, y)$$

We say that X and Y are *jointly continuous* if there exists a function $f(x, y)$, defined for all real x and y, having the property that for all sets A and B of real numbers

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x, y)\, dx\, dy$$

The function $f(x, y)$ is called the *joint probability density function* of X and Y. The probability density of X can be obtained from a knowledge of $f(x, y)$ by the following reasoning:

$$\begin{aligned} P\{X \in A\} &= P\{X \in A, Y \in (-\infty, \infty)\} \\ &= \int_{-\infty}^{\infty} \int_A f(x, y)\, dx\, dy \\ &= \int_A f_X(x)\, dx \end{aligned}$$

where

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

is thus the probability density function of X. Similarly, the probability density function of Y is given by

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

Because

$$F(a, b) = P(X \le a, Y \le b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y)\, dy\, dx$$

differentiation yields

$$\frac{d^2}{da\, db} F(a, b) = f(a, b)$$

Thus, as in the single variable case, differentiating the probability distribution function gives the probability density function.

A variation of Proposition 2.1 states that if X and Y are random variables and g is a function of two variables, then

$$\begin{aligned} E[g(X, Y)] &= \sum_y \sum_x g(x, y) p(x, y) && \text{in the discrete case} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy && \text{in the continuous case} \end{aligned}$$

For example, if $g(X, Y) = X + Y$, then, in the continuous case,

$$\begin{aligned} E[X + Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y)\, dx\, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y)\, dx\, dy \\ &= E[X] + E[Y] \end{aligned}$$

where the first integral is evaluated by using the variation of Proposition 2.1 with $g(x, y) = x$, and the second with $g(x, y) = y$.

The same result holds in the discrete case and, combined with the corollary in Section 2.4.3, yields that for any constants a, b

$$E[aX + bY] = aE[X] + bE[Y] \tag{2.10}$$

Joint probability distributions may also be defined for n random variables. The details are exactly the same as when $n = 2$ and are left as an exercise. The corresponding result to Equation (2.10) states that if $X_1, X_2, \ldots, X_n$ are n random variables, then for any n constants $a_1, a_2, \ldots, a_n$,

$$E[a_1 X_1 + a_2 X_2 + \cdots + a_n X_n] = a_1 E[X_1] + a_2 E[X_2] + \cdots + a_n E[X_n] \tag{2.11}$$

Example 2.29 Calculate the expected sum obtained when three fair dice are rolled.

Solution: Let X denote the sum obtained. Then $X = X_1 + X_2 + X_3$ where X_i represents the value of the ith die. Thus,

$$E[X] = E[X_1] + E[X_2] + E[X_3] = 3\left(\tfrac{7}{2}\right) = \tfrac{21}{2}$$ ■

Example 2.30 As another example of the usefulness of Equation (2.11), let us use it to obtain the expectation of a binomial random variable having parameters n and p. Recalling that such a random variable X represents the number of successes in n trials when each trial has probability p of being a success, we have

$$X = X_1 + X_2 + \cdots + X_n$$

where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{if the } i\text{th trial is a failure} \end{cases}$$

Hence, X_i is a Bernoulli random variable having expectation $E[X_i] = 1(p) + 0(1-p) = p$. Thus,

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = np$$

This derivation should be compared with the one presented in Example 2.17. ■

Example 2.31 At a party N men throw their hats into the center of a room. The hats are mixed up and each man randomly selects one. Find the expected number of men who select their own hats.

Solution: Letting X denote the number of men that select their own hats, we can best compute $E[X]$ by noting that

$$X = X_1 + X_2 + \cdots + X_N$$

where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th man selects his own hat} \\ 0, & \text{otherwise} \end{cases}$$

Now, because the ith man is equally likely to select any of the N hats, it follows that

$$P\{X_i = 1\} = P\{i\text{th man selects his own hat}\} = \frac{1}{N}$$

and so

$$E[X_i] = 1P\{X_i = 1\} + 0P\{X_i = 0\} = \frac{1}{N}$$

Hence, from Equation (2.11) we obtain

$$E[X] = E[X_1] + \cdots + E[X_N] = \left(\frac{1}{N}\right) N = 1$$

Hence, no matter how many people are at the party, on the average exactly one of the men will select his own hat. ■

Example 2.32 Suppose there are 25 different types of coupons and suppose that each time one obtains a coupon, it is equally likely to be any one of the 25 types. Compute the expected number of different types that are contained in a set of 10 coupons.

Solution: Let X denote the number of different types in the set of 10 coupons. We compute $E[X]$ by using the representation

$$X = X_1 + \cdots + X_{25}$$

where

$$X_i = \begin{cases} 1, & \text{if at least one type } i \text{ coupon is in the set of 10} \\ 0, & \text{otherwise} \end{cases}$$

Now,

$$\begin{aligned} E[X_i] &= P\{X_i = 1\} \\ &= P\{\text{at least one type } i \text{ coupon is in the set of 10}\} \\ &= 1 - P\{\text{no type } i \text{ coupons are in the set of 10}\} \\ &= 1 - \left(\tfrac{24}{25}\right)^{10} \end{aligned}$$

when the last equality follows since each of the 10 coupons will (independently) not be a type i with probability $\frac{24}{25}$. Hence,

$$E[X] = E[X_1] + \cdots + E[X_{25}] = 25\left[1 - \left(\tfrac{24}{25}\right)^{10}\right]$$ ■

2.5.2 Independent Random Variables

The random variables X and Y are said to be *independent* if, for all a, b,

$$P\{X \le a, Y \le b\} = P\{X \le a\}P\{Y \le b\} \tag{2.12}$$

In other words, X and Y are independent if, for all a and b, the events $E_a = \{X \le a\}$ and $F_b = \{Y \le b\}$ are independent.

In terms of the joint distribution function F of X and Y, we have that X and Y are independent if

$$F(a, b) = F_X(a)F_Y(b) \qquad \text{for all } a, b$$

When X and Y are discrete, the condition of independence reduces to

$$p(x, y) = p_X(x)p_Y(y) \tag{2.13}$$

while if X and Y are jointly continuous, independence reduces to

$$f(x, y) = f_X(x)f_Y(y) \tag{2.14}$$

To prove this statement, consider first the discrete version, and suppose that the joint probability mass function $p(x, y)$ satisfies Equation (2.13). Then

$$\begin{aligned} P\{X \le a, Y \le b\} &= \sum_{y \le b} \sum_{x \le a} p(x, y) \\ &= \sum_{y \le b} \sum_{x \le a} p_X(x) p_Y(y) \\ &= \sum_{y \le b} p_Y(y) \sum_{x \le a} p_X(x) \\ &= P\{Y \le b\} P\{X \le a\} \end{aligned}$$

and so X and Y are independent. That Equation (2.14) implies independence in the continuous case is proven in the same manner and is left as an exercise.

An important result concerning independence is the following.

Proposition 2.3 If X and Y are independent, then for any functions h and g

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof. Suppose that X and Y are jointly continuous. Then

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) f(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) f_X(x) f_Y(y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} h(y) f_Y(y)\, dy \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \\ &= E[h(Y)]E[g(X)] \end{aligned}$$

The proof in the discrete case is similar. ■

2.5.3 Covariance and Variance of Sums of Random Variables

The covariance of any two random variables X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined by

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - YE[X] - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Note that if X and Y are independent, then by Proposition 2.3 it follows that $\mathrm{Cov}(X, Y) = 0$.

Let us consider now the special case where X and Y are indicator variables for whether or not the events A and B occur. That is, for events A and B, define

$$X = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise,} \end{cases} \qquad Y = \begin{cases} 1, & \text{if } B \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

and, because XY will equal 1 or 0 depending on whether or not both X and Y equal 1, we see that

$$\text{Cov}(X, Y) = P\{X = 1, Y = 1\} - P\{X = 1\}P\{Y = 1\}$$

From this we see that

$$\begin{aligned} \text{Cov}(X, Y) > 0 &\Leftrightarrow P\{X = 1, Y = 1\} > P\{X = 1\}P\{Y = 1\} \\ &\Leftrightarrow \frac{P\{X = 1, Y = 1\}}{P\{X = 1\}} > P\{Y = 1\} \\ &\Leftrightarrow P\{Y = 1|X = 1\} > P\{Y = 1\} \end{aligned}$$

That is, the covariance of X and Y is positive if the outcome $X = 1$ makes it more likely that $Y = 1$ (which, as is easily seen by symmetry, also implies the reverse).

In general it can be shown that a positive value of $\text{Cov}(X, Y)$ is an indication that Y tends to increase as X does, whereas a negative value indicates that Y tends to decrease as X increases.

Example 2.33 The joint density function of X, Y is

$$f(x, y) = \frac{1}{y} e^{-(y+x/y)}, \quad 0 < x, y < \infty$$

(a) Verify that the preceding is a joint density function.
(b) Find $\text{Cov}(X, Y)$.

Solution: To show that $f(x, y)$ is a joint density function we need to show it is nonnegative, which is immediate, and that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx = 1$. We prove the latter as follows:

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx &= \int_0^{\infty} \int_0^{\infty} \frac{1}{y} e^{-(y+x/y)} dy\, dx \\ &= \int_0^{\infty} e^{-y} \int_0^{\infty} \frac{1}{y} e^{-x/y} dx\, dy \\ &= \int_0^{\infty} e^{-y} dy \\ &= 1 \end{aligned}$$

To obtain $\text{Cov}(X, Y)$, note that the density funtion of Y is

$$f_Y(y) = e^{-y} \int_0^\infty \frac{1}{y} e^{-x/y} dx = e^{-y}$$

Thus, Y is an exponential random variable with parameter 1, showing (see Example 2.21) that

$$E[Y] = 1$$

We compute $E[X]$ and $E[XY]$ as follows:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y)\, dy\, dx \\ &= \int_0^\infty e^{-y} \int_0^\infty \frac{x}{y} e^{-x/y} dx\, dy \end{aligned}$$

Now, $\int_0^\infty \frac{x}{y} e^{-x/y} dx$ is the expected value of an exponential random variable with parameter $1/y$, and thus is equal to y. Consequently,

$$E[X] = \int_0^\infty y e^{-y} dy = 1$$

Also

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f(x, y)\, dy\, dx \\ &= \int_0^\infty y e^{-y} \int_0^\infty \frac{x}{y} e^{-x/y} dx\, dy \\ &= \int_0^\infty y^2 e^{-y} dy \end{aligned}$$

Integration by parts ($dv = e^{-y} dy$, $u = y^2$) gives

$$E[XY] = \int_0^\infty y^2 e^{-y} dy = -y^2 e^{-y}|_0^\infty + \int_0^\infty 2y e^{-y} dy = 2E[Y] = 2$$

Consequently,

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 1$$ ■

The following are important properties of covariance.

Properties of Covariance

For any random variables X, Y, Z and constant c,

1. $\text{Cov}(X, X) = \text{Var}(X)$,
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$,

3. $\text{Cov}(cX, Y) = c\,\text{Cov}(X, Y)$,
4. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$.

Whereas the first three properties are immediate, the final one is easily proven as follows:

$$\begin{aligned}\text{Cov}(X, Y + Z) &= E[X(Y + Z)] - E[X]E[Y + Z] \\ &= E[XY] - E[X]E[Y] + E[XZ] - E[X]E[Z] \\ &= \text{Cov}(X, Y) + \text{Cov}(X, Z)\end{aligned}$$

The fourth property listed easily generalizes to give the following result:

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j) \tag{2.15}$$

A useful expression for the variance of the sum of random variables can be obtained from Equation (2.15) as follows:

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j<i} \text{Cov}(X_i, X_j) \qquad (2.16)\end{aligned}$$

If $X_i, i = 1, \ldots, n$ are independent random variables, then Equation (2.16) reduces to

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

Definition 2.1 If $X_1, \ldots, X_n$ are independent and identically distributed, then the random variable $\bar{X} = \sum_{i=1}^{n} X_i/n$ is called the *sample mean*.

The following proposition shows that the covariance between the sample mean and a deviation from that sample mean is zero. It will be needed in Section 2.6.1.

Proposition 2.4 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed with expected value μ and variance σ^2. Then,

(a) $E[\bar{X}] = \mu$.
(b) $\text{Var}(\bar{X}) = \sigma^2/n$.
(c) $\text{Cov}(\bar{X}, X_i - \bar{X}) = 0, i = 1, \ldots, n$.

Proof. Parts (a) and (b) are easily established as follows:

$$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu,$$

$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

To establish part (c) we reason as follows:

$$\begin{aligned}
\mathrm{Cov}(\bar{X}, X_i - \bar{X}) &= \mathrm{Cov}(\bar{X}, X_i) - \mathrm{Cov}(\bar{X}, \bar{X}) \\
&= \frac{1}{n}\mathrm{Cov}\left(X_i + \sum_{j\neq i} X_j, X_i\right) - \mathrm{Var}(\bar{X}) \\
&= \frac{1}{n}\mathrm{Cov}(X_i, X_i) + \frac{1}{n}\mathrm{Cov}\left(\sum_{j\neq i} X_j, X_i\right) - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0
\end{aligned}$$

where the final equality used the fact that X_i and $\sum_{j\neq i} X_j$ are independent and thus have covariance 0. ■

Equation (2.16) is often useful when computing variances.

Example 2.34 (Variance of a Binomial Random Variable) Compute the variance of a binomial random variable X with parameters n and p.

Solution: Since such a random variable represents the number of successes in n independent trials when each trial has a common probability p of being a success, we may write

$$X = X_1 + \cdots + X_n$$

where the X_i are independent Bernoulli random variables such that

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{otherwise} \end{cases}$$

Hence, from Equation (2.16) we obtain

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$$

But

$$\begin{aligned}
\mathrm{Var}(X_i) &= E[X_i^2] - (E[X_i])^2 \\
&= E[X_i] - (E[X_i])^2 \qquad \text{since } X_i^2 = X_i \\
&= p - p^2
\end{aligned}$$

and thus

$$\mathrm{Var}(X) = np(1-p) \qquad \blacksquare$$

Example 2.35 (Sampling from a Finite Population: The Hypergeometric) Consider a population of N individuals, some of whom are in favor of a certain proposition. In particular suppose that Np of them are in favor and $N - Np$ are opposed, where p is assumed to be unknown. We are interested in estimating p, the fraction of the population that is for the proposition, by randomly choosing and then determining the positions of n members of the population.

In such situations as described in the preceding, it is common to use the fraction of the sampled population that is in favor of the proposition as an estimator of p. Hence, if we let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th person chosen is in favor} \\ 0, & \text{otherwise} \end{cases}$$

then the usual estimator of p is $\sum_{i=1}^{n} X_i/n$. Let us now compute its mean and variance. Now,

$$\begin{aligned} E\left[\sum_{i=1}^{n} X_i\right] &= \sum_{1}^{n} E[X_i] \\ &= np \end{aligned}$$

where the final equality follows since the ith person chosen is equally likely to be any of the N individuals in the population and so has probability Np/N of being in favor.

$$\mathrm{Var}\left(\sum_{1}^{n} X_i\right) = \sum_{1}^{n} \mathrm{Var}(X_i) + 2\sum\sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

Now, since X_i is a Bernoulli random variable with mean p, it follows that

$$\mathrm{Var}(X_i) = p(1-p)$$

Also, for $i \neq j$,

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] \\ &= P\{X_i = 1, X_j = 1\} - p^2 \\ &= P\{X_i = 1\}P\{X_j = 1 | X_i = 1\} - p^2 \\ &= \frac{Np}{N}\frac{(Np-1)}{N-1} - p^2 \end{aligned}$$

where the last equality follows since if the ith person to be chosen is in favor, then the jth person chosen is equally likely to be any of the other $N-1$ of which $Np-1$ are in favor. Thus, we see that

$$\begin{aligned} \mathrm{Var}\left(\sum_{1}^{n} X_i\right) &= np(1-p) + 2\binom{n}{2}\left[\frac{p(Np-1)}{N-1} - p^2\right] \\ &= np(1-p) - \frac{n(n-1)p(1-p)}{N-1} \end{aligned}$$

and so the mean and variance of our estimator are given by

$$E\left[\sum_1^n \frac{X_i}{n}\right] = p,$$

$$\operatorname{Var}\left[\sum_1^n \frac{X_i}{n}\right] = \frac{p(1-p)}{n} - \frac{(n-1)p(1-p)}{n(N-1)}$$

Some remarks are in order: As the mean of the estimator is the unknown value p, we would like its variance to be as small as possible (why is this?), and we see by the preceding that, as a function of the population size N, the variance increases as N increases. The limiting value, as $N \to \infty$, of the variance is $p(1-p)/n$, which is not surprising since for N large each of the X_i will be (approximately) independent random variables, and thus $\sum_1^n X_i$ will have an (approximately) binomial distribution with parameters n and p.

The random variable $\sum_1^n X_i$ can be thought of as representing the number of white balls obtained when n balls are randomly selected from a population consisting of Np white and $N - Np$ black balls. (Identify a person who favors the proposition with a white ball and one against with a black ball.) Such a random variable is called *hypergeometric* and has a probability mass function given by

$$P\left\{\sum_1^n X_i = k\right\} = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}} \qquad \blacksquare$$

It is often important to be able to calculate the distribution of $X + Y$ from the distributions of X and Y when X and Y are independent. Suppose first that X and Y are continuous, X having probability density f and Y having probability density g. Then, letting $F_{X+Y}(a)$ be the cumulative distribution function of $X + Y$, we have

$$\begin{aligned}
F_{X+Y}(a) &= P\{X + Y \le a\} \\
&= \iint_{x+y\le a} f(x)g(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f(x)g(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-y} f(x)\,dx\right) g(y)\,dy \\
&= \int_{-\infty}^{\infty} F_X(a-y)g(y)\,dy \qquad (2.17)
\end{aligned}$$

The cumulative distribution function F_{X+Y} is called the *convolution* of the distributions F_X and F_Y (the cumulative distribution functions of X and Y, respectively).

By differentiating Equation (2.17), we obtain that the probability density function $f_{X+Y}(a)$ of $X + Y$ is given by

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da}\int_{-\infty}^{\infty} F_X(a-y)g(y)\,dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da}(F_X(a-y))g(y)\,dy \\ &= \int_{-\infty}^{\infty} f(a-y)g(y)\,dy \end{aligned} \tag{2.18}$$

Example 2.36 (Sum of Two Independent Uniform Random Variables) If X and Y are independent random variables both uniformly distributed on (0, 1), then calculate the probability density of $X + Y$.

Solution: From Equation (2.18), since

$$f(a) = g(a) = \begin{cases} 1, & 0 < a < 1 \\ 0, & \text{otherwise} \end{cases}$$

we obtain

$$f_{X+Y}(a) = \int_0^1 f(a-y)\,dy$$

For $0 \le a \le 1$, this yields

$$f_{X+Y}(a) = \int_0^a dy = a$$

For $1 < a < 2$, we get

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2 - a$$

Hence,

$$f_{X+Y}(a) = \begin{cases} a, & 0 \le a \le 1 \\ 2-a, & 1 < a < 2 \\ 0, & \text{otherwise} \end{cases}$$

■

Rather than deriving a general expression for the distribution of $X+Y$ in the discrete case, we shall consider an example.

Example 2.37 (Sums of Independent Poisson Random Variables) Let X and Y be independent Poisson random variables with respective means λ_1 and λ_2. Calculate the distribution of $X + Y$.

Solution: Since the event $\{X + Y = n\}$ may be written as the union of the disjoint events $\{X = k, Y = n - k\}$, $0 \le k \le n$, we have

$$\begin{aligned}
P\{X + Y = n\} &= \sum_{k=0}^{n} P\{X = k, Y = n - k\} \\
&= \sum_{k=0}^{n} P\{X = k\}P\{Y = n - k\} \\
&= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n
\end{aligned}$$

In words, $X_1 + X_2$ has a Poisson distribution with mean $\lambda_1 + \lambda_2$. ■

The concept of independence may, of course, be defined for more than two random variables. In general, the n random variables $X_1, X_2, \ldots, X_n$ are said to be independent if, for all values $a_1, a_2, \ldots, a_n$,

$$\begin{aligned}
&P\{X_1 \le a_1, X_2 \le a_2, \ldots, X_n \le a_n\} \\
&\quad = P\{X_1 \le a_1\}P\{X_2 \le a_2\} \cdots P\{X_n \le a_n\}
\end{aligned}$$

Example 2.38 Let $X_1, \ldots, X_n$ be independent and identically distributed continuous random variables with probability distribution F and density function $F' = f$. If we let $X_{(i)}$ denote the ith smallest of these random variables, then $X_{(1)}, \ldots, X_{(n)}$ are called the *order statistics*. To obtain the distribution of $X_{(i)}$, note that $X_{(i)}$ will be less than or equal to x if and only if at least i of the n random variables $X_1, \ldots, X_n$ are less than or equal to x. Hence,

$$P\{X_{(i)} \le x\} = \sum_{k=i}^{n} \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}$$

Differentiation yields that the density function of $X_{(i)}$ is as follows:

$$
\begin{aligned}
f_{X_{(i)}}(x) &= f(x)\sum_{k=i}^{n}\binom{n}{k}k(F(x))^{k-1}(1-F(x))^{n-k} \\
&\quad -f(x)\sum_{k=i}^{n}\binom{n}{k}(n-k)(F(x))^{k}(1-F(x))^{n-k-1} \\
&= f(x)\sum_{k=i}^{n}\frac{n!}{(n-k)!(k-1)!}(F(x))^{k-1}(1-F(x))^{n-k} \\
&\quad -f(x)\sum_{k=i}^{n-1}\frac{n!}{(n-k-1)!k!}(F(x))^{k}(1-F(x))^{n-k-1} \\
&= f(x)\sum_{k=i}^{n}\frac{n!}{(n-k)!(k-1)!}(F(x))^{k-1}(1-F(x))^{n-k} \\
&\quad -f(x)\sum_{j=i+1}^{n}\frac{n!}{(n-j)!(j-1)!}(F(x))^{j-1}(1-F(x))^{n-j} \\
&= \frac{n!}{(n-i)!(i-1)!}f(x)(F(x))^{i-1}(1-F(x))^{n-i}
\end{aligned}
$$

The preceding density is quite intuitive, since in order for $X_{(i)}$ to equal x, $i-1$ of the n values $X_1, \ldots, X_n$ must be less than x; $n-i$ of them must be greater than x; and one must be equal to x. Now, the probability density that every member of a specified set of $i-1$ of the X_j is less than x, every member of another specified set of $n-i$ is greater than x, and the remaining value is equal to x is $(F(x))^{i-1}(1-F(x))^{n-i}f(x)$. Therefore, since there are $n!/[(i-1)!(n-i)!]$ different partitions of the n random variables into the three groups, we obtain the preceding density function. ■

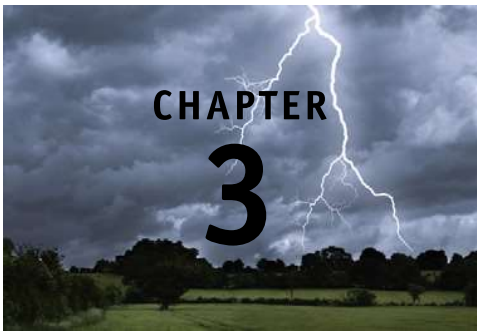
2.5.4 Joint Probability Distribution of Functions of Random Variables

Let X_1 and X_2 be jointly continuous random variables with joint probability density function $f(x_1, x_2)$. It is sometimes necessary to obtain the joint distribution of the random variables Y_1 and Y_2 that arise as functions of X_1 and X_2. Specifically, suppose that $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ for some functions g_1 and g_2.

Assume that the functions g_1 and g_2 satisfy the following conditions:

1. The equations $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ can be uniquely solved for x_1 and x_2 in terms of y_1 and y_2 with solutions given by, say, $x_1 = h_1(y_1, y_2)$, $x_2 = h_2(y_1, y_2)$.

Conditional Probability and Conditional Expectation



3.1 Introduction

One of the most useful concepts in probability theory is that of conditional probability and conditional expectation. The reason is twofold. First, in practice, we are often interested in calculating probabilities and expectations when some partial information is available; hence, the desired probabilities and expectations are conditional ones. Secondly, in calculating a desired probability or expectation it is often extremely useful to first "condition" on some appropriate random variable.

3.2 The Discrete Case

Recall that for any two events E and F, the conditional probability of E given F is defined, as long as $P(F) > 0$, by

$$P(E|F) = \frac{P(EF)}{P(F)}$$

Hence, if X and Y are discrete random variables, then it is natural to define the *conditional probability mass function* of X given that $Y = y$, by

$$\begin{aligned} p_{X|Y}(x|y) &= P\{X = x|Y = y\} \\ &= \frac{P\{X = x, Y = y\}}{P\{Y = y\}} \\ &= \frac{p(x, y)}{p_Y(y)} \end{aligned}$$

Introduction to Probability Models, Eleventh Edition. http://dx.doi.org/10.1016/B978-0-12-407948-9.00003-7
© 2014 Elsevier Inc. All rights reserved.

for all values of y such that $P\{Y = y\} > 0$. Similarly, the conditional probability distribution function of X given that $Y = y$ is defined, for all y such that $P\{Y = y\} > 0$, by

$$\begin{aligned} F_{X|Y}(x|y) &= P\{X \leqslant x|Y = y\} \\ &= \sum_{a \leqslant x} p_{X|Y}(a|y) \end{aligned}$$

Finally, the conditional expectation of X given that $Y = y$ is defined by

$$\begin{aligned} E[X|Y = y] &= \sum_x x P\{X = x|Y = y\} \\ &= \sum_x x p_{X|Y}(x|y) \end{aligned}$$

In other words, the definitions are exactly as before with the exception that everything is now conditional on the event that $Y = y$. If X is independent of Y, then the conditional mass function, distribution, and expectation are the same as the unconditional ones. This follows, since if X is independent of Y, then

$$\begin{aligned} p_{X|Y}(x|y) &= P\{X = x|Y = y\} \\ &= P\{X = x\} \end{aligned}$$

Example 3.1 Suppose that $p(x, y)$, the joint probability mass function of X and Y, is given by

$$p(1, 1) = 0.5, \quad p(1, 2) = 0.1, \quad p(2, 1) = 0.1, \quad p(2, 2) = 0.3$$

Calculate the conditional probability mass function of X given that $Y = 1$.

Solution: We first note that

$$p_Y(1) = \sum_x p(x, 1) = p(1, 1) + p(2, 1) = 0.6$$

Hence,

$$\begin{aligned} p_{X|Y}(1|1) &= P\{X = 1|Y = 1\} \\ &= \frac{P\{X = 1, Y = 1\}}{P\{Y = 1\}} \\ &= \frac{p(1, 1)}{p_Y(1)} \\ &= \frac{5}{6} \end{aligned}$$

Similarly,

$$p_{X|Y}(2|1) = \frac{p(2, 1)}{p_Y(1)} = \frac{1}{6}$$

■

Example 3.2 If X_1 and X_2 are independent binomial random variables with respective parameters (n_1, p) and (n_2, p), calculate the conditional probability mass function of X_1 given that $X_1 + X_2 = m$.

Solution: With $q = 1 - p$,

$$\begin{aligned} P\{X_1 = k | X_1 + X_2 = m\} &= \frac{P\{X_1 = k, X_1 + X_2 = m\}}{P\{X_1 + X_2 = m\}} \\ &= \frac{P\{X_1 = k, X_2 = m - k\}}{P\{X_1 + X_2 = m\}} \\ &= \frac{P\{X_1 = k\}P\{X_2 = m - k\}}{P\{X_1 + X_2 = m\}} \\ &= \frac{\binom{n_1}{k} p^k q^{n_1 - k} \binom{n_2}{m-k} p^{m-k} q^{n_2 - m + k}}{\binom{n_1 + n_2}{m} p^m q^{n_1 + n_2 - m}} \end{aligned}$$

where we have used that $X_1 + X_2$ is a binomial random variable with parameters $(n_1 + n_2, p)$ (see Example 2.44). Thus, the conditional probability mass function of X_1, given that $X_1 + X_2 = m$, is

$$P\{X_1 = k | X_1 + X_2 = m\} = \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n_1 + n_2}{m}} \tag{3.1}$$

The distribution given by Equation (3.1), first seen in Example 2.35, is known as the *hypergeometric* distribution. It is the distribution of the number of blue balls that are chosen when a sample of m balls is randomly chosen from an urn that contains n_1 blue and n_2 red balls. (To intuitively see why the conditional distribution is hypergeometric, consider $n_1 + n_2$ independent trials that each result in a success with probability p; let X_1 represent the number of successes in the first n_1 trials and let X_2 represent the number of successes in the final n_2 trials. Because all trials have the same probability of being a success, each of the $\binom{n_1+n_2}{m}$ subsets of m trials is equally likely to be the success trials; thus, the number of the m success trials that are among the first n_1 trials is a hypergeometric random variable.) ■

Example 3.3 If X and Y are independent Poisson random variables with respective means λ_1 and λ_2, calculate the conditional expected value of X given that $X + Y = n$.

Solution: Let us first calculate the conditional probability mass function of X given that $X + Y = n$. We obtain

$$\begin{aligned} P\{X = k|X + Y = n\} &= \frac{P\{X = k, X + Y = n\}}{P\{X + Y = n\}} \\ &= \frac{P\{X = k, Y = n - k\}}{P\{X + Y = n\}} \\ &= \frac{P\{X = k\}P\{Y = n - k\}}{P\{X + Y = n\}} \end{aligned}$$

where the last equality follows from the assumed independence of X and Y. Recalling (see Example 2.37) that $X + Y$ has a Poisson distribution with mean $\lambda_1 + \lambda_2$, the preceding equation equals

$$\begin{aligned} P\{X = k|X + Y = n\} &= \frac{e^{-\lambda_1}\lambda_1^k}{k!}\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}\left[\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}\right]^{-1} \\ &= \frac{n!}{(n-k)!k!}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \end{aligned}$$

In other words, the conditional distribution of X given that $X+Y = n$ is the binomial distribution with parameters n and $\lambda_1/(\lambda_1 + \lambda_2)$. Hence,

$$E\{X|X + Y = n\} = n\frac{\lambda_1}{\lambda_1 + \lambda_2}$$ ■

Conditional expectations possess all of the properties of ordinary expectations. For example such identities such as

$$E\left[\sum_{i=1}^{n} X_i|Y = y\right] = \sum_{i=1}^{n} E[X_i|Y = y]$$

$$E[h(X)|Y = y] = \sum_x h(x)P(X = x|Y = y)$$

remain valid.

Example 3.4 There are n components. On a rainy day, component i will function with probability p_i; on a nonrainy day, component i will function with probability q_i, for $i = 1, \ldots, n$. It will rain tomorrow with probability α. Calculate the conditional expected number of components that function tomorrow, given that it rains.

Solution: Let

$$X_i = \begin{cases} 1, & \text{if component } i \text{ functions tomorrow} \\ 0, & \text{otherwise} \end{cases}$$

Then, with Y defined to equal 1 if it rains tomorrow, and 0 otherwise, the desired conditional expectation is obtained as follows.

$$\begin{aligned} E\left[\sum_{t=1}^{n} X_i | Y = 1\right] &= \sum_{i=1}^{n} E[X_i | Y = 1] \\ &= \sum_{i=1}^{n} p_i \end{aligned}$$

■

3.3 The Continuous Case

If X and Y have a joint probability density function $f(x, y)$, then the *conditional probability density function* of X, given that $Y = y$, is defined for all values of y such that $f_Y(y) > 0$, by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

To motivate this definition, multiply the left side by dx and the right side by $(dx\, dy)/dy$ to get

$$\begin{aligned} f_{X|Y}(x|y)\, dx &= \frac{f(x, y)\, dx\, dy}{f_Y(y)\, dy} \\ &\approx \frac{P\{x \leqslant X \leqslant x + dx, y \leqslant Y \leqslant y + dy\}}{P\{y \leqslant Y \leqslant y + dy\}} \\ &= P\{x \leqslant X \leqslant x + dx | y \leqslant Y \leqslant y + dy\} \end{aligned}$$

In other words, for small values dx and dy, $f_{X|Y}(x|y)\, dx$ is approximately the conditional probability that X is between x and $x + dx$ given that Y is between y and $y + dy$.

The *conditional expectation* of X, given that $Y = y$, is defined for all values of y such that $f_Y(y) > 0$, by

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$

Example 3.5 Suppose the joint density of X and Y is given by

$$f(x, y) = \begin{cases} 6xy(2 - x - y), & 0 < x < 1, 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Compute the conditional expectation of X given that $Y = y$, where $0 < y < 1$.

Solution: We first compute the conditional density

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{6xy(2-x-y)}{\int_0^1 6xy(2-x-y)\,dx} \\ &= \frac{6xy(2-x-y)}{y(4-3y)} \\ &= \frac{6x(2-x-y)}{4-3y} \end{aligned}$$

Hence,

$$\begin{aligned} E[X|Y=y] &= \int_0^1 \frac{6x^2(2-x-y)\,dx}{4-3y} \\ &= \frac{(2-y)2-\frac{6}{4}}{4-3y} \\ &= \frac{5-4y}{8-6y} \end{aligned}$$

■

Example 3.6 (The t-Distribution) If Z and Y are independent, with Z having a standard normal distribution and Y having a chi-squared distribution with n degrees of freedom, then the random variable T defined by

$$T = \frac{Z}{\sqrt{Y/n}} = \sqrt{n}\frac{Z}{\sqrt{Y}}$$

is said to be a *t-random variable* with n degrees of freedom. To compute its density function, we first derive the conditional density of T given that $Y = y$. Because Z and Y are independent, the conditional distribution of T given that $Y = y$ is the distribution of $\sqrt{n/y}Z$, which is normal with mean 0 and variance n/y. Hence, the conditional density function of T given that $Y = y$ is

$$f_{T|Y}(t|y) = \frac{1}{\sqrt{2\pi n/y}}e^{-t^2y/2n} = \frac{y^{1/2}}{\sqrt{2\pi n}}e^{-t^2y/2n}, \quad -\infty < t < \infty$$

Using the preceding, along with the following formula for the chi-squared density derived in Exercise 87 of Chapter 2,

$$f_Y(y) = \frac{e^{-y/2}y^{n/2-1}}{2^{n/2}\Gamma(n/2)}, \quad y > 0$$

we obtain the density function of T:

$$f_T(t) = \int_0^\infty f_{T,Y}(t,y)\,dy = \int_0^\infty f_{T|Y}(t|y)f_Y(y)\,dy$$

With

$$K = \frac{1}{\sqrt{\pi n}\, 2^{(n+1)/2}\Gamma(n/2)}, \quad c = \frac{t^2+n}{2n} = \frac{1}{2}\left(1+\frac{t^2}{n}\right)$$

this yields

$$\begin{aligned} f_T(t) &= \frac{1}{K}\int_0^\infty e^{-cy} y^{(n-1)/2}\, dy \\ &= \frac{c^{-(n+1)/2}}{K}\int_0^\infty e^{-x} x^{(n-1)/2}\, dx \quad \text{(by letting } x = cy) \\ &= \frac{c^{-(n+1)/2}}{K}\Gamma\left(\frac{n+1}{2}\right) \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty \end{aligned}$$

■

Example 3.7 The joint density of X and Y is given by

$$f(x, y) = \begin{cases} \frac{1}{2} y e^{-xy}, & 0 < x < \infty, 0 < y < 2 \\ 0, & \text{otherwise} \end{cases}$$

What is $E[e^{X/2}|Y = 1]$?

Solution: The conditional density of X, given that $Y = 1$, is given by

$$\begin{aligned} f_{X|Y}(x|1) &= \frac{f(x, 1)}{f_Y(1)} \\ &= \frac{\frac{1}{2}e^{-x}}{\int_0^\infty \frac{1}{2}e^{-x}\, dx} = e^{-x} \end{aligned}$$

Hence, by Proposition 2.1,

$$\begin{aligned} E\left[e^{X/2}|Y = 1\right] &= \int_0^\infty e^{x/2} f_{X|Y}(x|1)\, dx \\ &= \int_0^\infty e^{x/2} e^{-x}\, dx \\ &= 2 \end{aligned}$$

■

Example 3.8 Let X_1 and X_2 be independent exponential random variables with rates μ_1 and μ_2. Find the conditional density of X_1 given that $X_1 + X_2 = t$.

Solution: To begin, let us first note that if $f(x, y)$ is the joint density of X, Y, then the joint density of X and $X + Y$ is

$$f_{X,X+Y}(x, t) = f(x, t - x)$$

which is easily seen by noting that the Jacobian of the transformation

$$g_1(x, y) = x, \quad g_2(x, y) = x + y$$

is equal to 1.
Applying the preceding to our example yields

$$\begin{aligned} f_{X_1|X_1+X_2}(x|t) &= \frac{f_{X_1,X_1+X_2}(x,t)}{f_{X_1+X_2}(t)} \\ &= \frac{\mu_1 e^{-\mu_1 x}\mu_2 e^{-\mu_2(t-x)}}{f_{X_1+X_2}(t)}, \quad 0 \leqslant x \leqslant t \\ &= Ce^{-(\mu_1-\mu_2)x}, \quad 0 \leqslant x \leqslant t \end{aligned}$$

where

$$C = \frac{\mu_1\mu_2 e^{-\mu_2 t}}{f_{X_1+X_2}(t)}$$

Now, if $\mu_1 = \mu_2$, then

$$f_{X_1|X_1+X_2}(x|t) = C, \quad 0 \leqslant x \leqslant t$$

yielding that $C = 1/t$ and that X_1 given $X_1 + X_2 = t$ is uniformly distributed on $(0, t)$. On the other hand, if $\mu_1 \neq \mu_2$, then we use

$$1 = \int_0^t f_{X_1|X_1+X_2}(x|t)dx = \frac{C}{\mu_1 - \mu_2}\left(1 - e^{-(\mu_1-\mu_2)t}\right)$$

to obtain

$$C = \frac{\mu_1 - \mu_2}{1 - e^{-(\mu_1-\mu_2)t}}$$

thus yielding the result:

$$f_{X_1|X_1+X_2}(x|t) = \frac{(\mu_1 - \mu_2)e^{-(\mu_1-\mu_2)x}}{1 - e^{-(\mu_1-\mu_2)t}}$$

An interesting byproduct of our analysis is that

$$f_{X_1+X_2}(t) = \frac{\mu_1\mu_2 e^{-\mu_2 t}}{C} = \begin{cases} \mu^2 t e^{-\mu t}, & \text{if } \mu_1 = \mu_2 = \mu \\ \frac{\mu_1\mu_2(e^{-\mu_2 t} - e^{-\mu_1 t})}{\mu_1 - \mu_2}, & \text{if } \mu_1 \neq \mu_2 \end{cases}$$ ■

3.4 Computing Expectations by Conditioning

Let us denote by $E[X|Y]$ that function of the random variable Y whose value at $Y = y$ is $E[X|Y = y]$. Note that $E[X|Y]$ is itself a random variable. An extremely important property of conditional expectation is that for all random variables X and Y

$$E[X] = E\big[E[X|Y]\big] \tag{3.2}$$

If Y is a discrete random variable, then Equation (3.2) states that

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\} \tag{3.2a}$$

while if Y is continuous with density $f_Y(y)$, then Equation (3.2) says that

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)\,dy \tag{3.2b}$$

We now give a proof of Equation (3.2) in the case where X and Y are both discrete random variables.

Proof of Equation (3.2) When X and Y Are Discrete. We must show that

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\} \tag{3.3}$$

Now, the right side of the preceding can be written

$$\begin{aligned}
\sum_y E[X|Y = y]P\{Y = y\} &= \sum_y \sum_x x P\{X = x|Y = y\}P\{Y = y\} \\
&= \sum_y \sum_x x \frac{P\{X = x, Y = y\}}{P\{Y = y\}} P\{Y = y\} \\
&= \sum_y \sum_x x P\{X = x, Y = y\} \\
&= \sum_x x \sum_y P\{X = x, Y = y\} \\
&= \sum_x x P\{X = x\} \\
&= E[X]
\end{aligned}$$

and the result is obtained. ■

One way to understand Equation (3.3) is to interpret it as follows. It states that to calculate $E[X]$ we may take a weighted average of the conditional expected value of X given that $Y = y$, each of the terms $E[X|Y = y]$ being weighted by the probability of the event on which it is conditioned.

The following examples will indicate the usefulness of Equation (3.2).

Example 3.9 Sam will read either one chapter of his probability book or one chapter of his history book. If the number of misprints in a chapter of his probability book is Poisson distributed with mean 2 and if the number of misprints in his history chapter is Poisson distributed with mean 5, then assuming Sam is equally likely to choose either book, what is the expected number of misprints that Sam will come across?

Solution: Letting X denote the number of misprints and letting

$$Y = \begin{cases} 1, & \text{if Sam chooses his history book} \\ 2, & \text{if Sam chooses his probability book} \end{cases}$$

then

$$\begin{aligned} E[X] &= E[X|Y=1]P\{Y=1\} + E[X|Y=2]P\{Y=2\} \\ &= 5\left(\tfrac{1}{2}\right) + 2\left(\tfrac{1}{2}\right) \\ &= \tfrac{7}{2} \end{aligned}$$

■

Example 3.10 (The Expectation of the Sum of a Random Number of Random Variables) Suppose that the expected number of accidents per week at an industrial plant is four. Suppose also that the numbers of workers injured in each accident are independent random variables with a common mean of 2. Assume also that the number of workers injured in each accident is independent of the number of accidents that occur. What is the expected number of injuries during a week?

Solution: Letting N denote the number of accidents and X_i the number injured in the ith accident, $i = 1, 2, \ldots$, then the total number of injuries can be expressed as $\sum_{i=1}^{N} X_i$. Now,

$$E\left[\sum_{1}^{N} X_i\right] = E\left[E\left[\sum_{1}^{N} X_i | N\right]\right]$$

But

$$\begin{aligned} E\left[\sum_{1}^{N} X_i | N = n\right] &= E\left[\sum_{1}^{n} X_i | N = n\right] \\ &= E\left[\sum_{1}^{n} X_i\right] \quad \text{by the independence of } X_i \text{ and } N \\ &= nE[X] \end{aligned}$$

which yields

$$E\left[\sum_{i=1}^{N} X_i | N\right] = NE[X]$$

and thus

$$E\left[\sum_{i=1}^{N} X_i\right] = E\big[NE[X]\big] = E[N]E[X]$$

Therefore, in our example, the expected number of injuries during a week equals $4 \times 2 = 8$. ■

The random variable $\sum_{i=1}^{N} X_i$, equal to the sum of a random number N of independent and identically distributed random variables that are also independent of N, is called a *compound random variable*. As just shown in Example 3.10, the expected value of a compound random variable is $E[X]E[N]$. Its variance will be derived in Example 3.19.

Example 3.11 (The Mean of a Geometric Distribution) A coin, having probability p of coming up heads, is to be successively flipped until the first head appears. What is the expected number of flips required?

Solution: Let N be the number of flips required, and let

$$Y = \begin{cases} 1, & \text{if the first flip results in a head} \\ 0, & \text{if the first flip results in a tail} \end{cases}$$

Now,

$$\begin{aligned} E[N] &= E[N|Y=1]P\{Y=1\} + E[N|Y=0]P\{Y=0\} \\ &= pE[N|Y=1] + (1-p)E[N|Y=0] \end{aligned} \tag{3.4}$$

However,

$$E[N|Y=1] = 1, \quad E[N|Y=0] = 1 + E[N] \tag{3.5}$$

To see why Equation (3.5) is true, consider $E[N|Y=1]$. Since $Y=1$, we know that the first flip resulted in heads and so, clearly, the expected number of flips required is 1. On the other hand if $Y=0$, then the first flip resulted in tails. However, since the successive flips are assumed independent, it follows that, after the first tail, the expected additional number of flips until the first head is just $E[N]$. Hence $E[N|Y=0] = 1 + E[N]$. Substituting Equation (3.5) into Equation (3.4) yields

$$E[N] = p + (1-p)(1 + E[N])$$

or

$$E[N] = 1/p$$ ■

Because the random variable N is a geometric random variable with probability mass function $p(n) = p(1-p)^{n-1}$, its expectation could easily have been computed from $E[N] = \sum_1^\infty np(n)$ without recourse to conditional expectation. However, if you attempt to obtain the solution to our next example without using conditional expectation, you will quickly learn what a useful technique "conditioning" can be.

Example 3.12 A miner is trapped in a mine containing three doors. The first door leads to a tunnel that takes him to safety after two hours of travel. The second door leads to a tunnel that returns him to the mine after three hours of travel. The third door leads to a tunnel that returns him to his mine after five hours. Assuming that the miner is at all times equally likely to choose any one of the doors, what is the expected length of time until the miner reaches safety?

Solution: Let X denote the time until the miner reaches safety, and let Y denote the door he initially chooses. Now,

$$\begin{aligned} E[X] &= E[X|Y=1]P\{Y=1\} + E[X|Y=2]P\{Y=2\} \\ &\quad + E[X|Y=3]P\{Y=3\} \\ &= \tfrac{1}{3}\big(E[X|Y=1] + E[X|Y=2] + E[X|Y=3]\big) \end{aligned}$$

However,

$$\begin{aligned} E[X|Y=1] &= 2, \\ E[X|Y=2] &= 3 + E[X], \\ E[X|Y=3] &= 5 + E[X], \end{aligned} \tag{3.6}$$

To understand why this is correct consider, for instance, $E[X|Y = 2]$, and reason as follows. If the miner chooses the second door, then he spends three hours in the tunnel and then returns to the mine. But once he returns to the mine the problem is as before, and hence his expected additional time until safety is just $E[X]$. Hence $E[X|Y = 2] = 3 + E[X]$. The argument behind the other equalities in Equation (3.6) is similar. Hence,

$$E[X] = \tfrac{1}{3}\big(2 + 3 + E[X] + 5 + E[X]\big) \quad \text{or} \quad E[X] = 10$$ ■

Example 3.13 (Multinomial Covariances) Consider n independent trials, each of which results in one of the outcomes $1, \ldots, r$, with respective probabilities $p_1, \ldots, p_r$, $\sum_{i=1}^{r} p_i = 1$. If we let N_i denote the number of trials that result in outcome i, then $(N_1, \ldots, N_r)$ is said to have a *multinomial distribution*. For $i \neq j$, let us compute

$$\text{Cov}(N_i, N_j) = E[N_i N_j] - E[N_i]E[N_j]$$

Because each trial independently results in outcome i with probability p_i, it follows that N_i is binomial with parameters (n, p_i), giving that $E[N_i]E[N_j] = n^2 p_i p_j$. To compute $E[N_i N_j]$, condition on N_i to obtain

$$\begin{aligned} E[N_i N_j] &= \sum_{k=0}^{n} E[N_i N_j | N_i = k] P(N_i = k) \\ &= \sum_{k=0}^{n} k E[N_j | N_i = k] P(N_j = k) \end{aligned}$$

Now, given that k of the n trials result in outcome i, each of the other $n - k$ trials independently results in outcome j with probability

$$P(j|\text{not } i) = \frac{p_j}{1 - p_i}$$

thus showing that the conditional distribution of N_j, given that $N_i = k$, is binomial with parameters $(n - k, \frac{p_j}{1-p_i})$. Using this yields

$$\begin{aligned} E[N_i N_j] &= \sum_{k=0}^{n} k(n-k) \frac{p_j}{1 - p_i} P(N_i = k) \\ &= \frac{p_j}{1 - p_i} \left(n \sum_{k=0}^{n} k P(N_i = k) - \sum_{k=0}^{n} k^2 P(N_i = k) \right) \\ &= \frac{p_j}{1 - p_i} (n E[N_i] - E[N_i^2]) \end{aligned}$$

Because N_i is binomial with parameters (n, p_i)

$$E[N_i^2] = \text{Var}(N_i) + (E[N_i])^2 = np_i(1 - p_i) + (np_i)^2$$

Hence,

$$\begin{aligned} E[N_i N_j] &= \frac{p_j}{1 - p_i}[n^2 p_i - np_i(1 - p_i) - n^2 p_i^2] \\ &= \frac{np_i p_j}{1 - p_i}[n - np_i - (1 - p_i)] \\ &= n(n - 1)p_i p_j \end{aligned}$$

which yields the result

$$\text{Cov}(N_i, N_j) = n(n - 1)p_i p_j - n^2 p_i p_j = -np_i p_j$$ ■

Example 3.14 (The Matching Rounds Problem) Suppose in Example 2.31 that those choosing their own hats depart, while the others (those without a match) put their selected hats in the center of the room, mix them up, and then reselect. Also, suppose that this process continues until each individual has his own hat.

(a) Find $E[R_n]$ where R_n is the number of rounds that are necessary when n individuals are initially present.
(b) Find $E[S_n]$ where S_n is the total number of selections made by the n individuals, $n \geqslant 2$.
(c) Find the expected number of false selections made by one of the n people, $n \geqslant 2$.

Solution: (a) It follows from the results of Example 2.31 that no matter how many people remain there will, on average, be one match per round. Hence, one might suggest that $E[R_n] = n$. This turns out to be true, and an induction proof will now be given. Because it is obvious that $E[R_1] = 1$, assume that $E[R_k] = k$ for $k = 1, \ldots, n - 1$. To compute $E[R_n]$, start by conditioning on X_n, the number of matches that occur in the first round. This gives

$$E[R_n] = \sum_{i=0}^{n} E[R_n | X_n = i] P\{X_n = i\}$$

Now, given a total of i matches in the initial round, the number of rounds needed will equal 1 plus the number of rounds that are required when $n - i$ persons are to be matched with their hats. Therefore,

$$\begin{aligned} E[R_n] &= \sum_{i=0}^{n} (1 + E[R_{n-i}]) P\{X_n = i\} \\ &= 1 + E[R_n] P\{X_n = 0\} + \sum_{i=1}^{n} E[R_{n-i}] P\{X_n = i\} \end{aligned}$$

$$
\begin{aligned}
&= 1 + E[R_n]P\{X_n = 0\} + \sum_{i=1}^{n} (n-i)P\{X_n = i\} \\
&\qquad \text{by the induction hypothesis} \\
&= 1 + E[R_n]P\{X_n = 0\} + n(1 - P\{X_n = 0\}) - E[X_n] \\
&= E[R_n]P\{X_n = 0\} + n(1 - P\{X_n = 0\})
\end{aligned}
$$

where the final equality used the result, established in Example 2.31, that $E[X_n] = 1$. Since the preceding equation implies that $E[R_n] = n$, the result is proven.
(b) For $n \geqslant 2$, conditioning on X_n, the number of matches in round 1, gives

$$
\begin{aligned}
E[S_n] &= \sum_{i=0}^{n} E[S_n | X_n = i]P\{X_n = i\} \\
&= \sum_{i=0}^{n} (n + E[S_{n-i}])P\{X_n = i\} \\
&= n + \sum_{i=0}^{n} E[S_{n-i}]P\{X_n = i\}
\end{aligned}
$$

where $E[S_0] = 0$. To solve the preceding equation, rewrite it as

$$E[S_n] = n + E[S_{n-X_n}]$$

Now, if there were *exactly* one match in each round, then it would take a total of $1 + 2 + \cdots + n = n(n+1)/2$ selections. Thus, let us try a solution of the form $E[S_n] = an + bn^2$. For the preceding equation to be satisfied by a solution of this type, for $n \geqslant 2$, we need

$$an + bn^2 = n + E\big[a(n - X_n) + b(n - X_n)^2\big]$$

or, equivalently,

$$an + bn^2 = n + a(n - E[X_n]) + b\big(n^2 - 2nE[X_n] + E\big[X_n^2\big]\big)$$

Now, using the results of Example 2.31 and Exercise 72 of Chapter 2 that $E[X_n] = \mathrm{Var}(X_n) = 1$, the preceding will be satisfied if

$$an + bn^2 = n + an - a + bn^2 - 2nb + 2b$$

and this will be valid provided that $b = 1/2$, $a = 1$. That is,

$$E[S_n] = n + n^2/2$$

satisfies the recursive equation for $E[S_n]$.

The formal proof that $E[S_n] = n + n^2/2$, $n \geqslant 2$, is obtained by induction on n. It is true when $n = 2$ (since, in this case, the number of selections is twice the number

of rounds and the number of rounds is a geometric random variable with parameter $p = 1/2$). Now, the recursion gives

$$E[S_n] = n + E[S_n]P\{X_n = 0\} + \sum_{i=1}^{n} E[S_{n-i}]P\{X_n = i\}$$

Hence, upon assuming that $E[S_0] = E[S_1] = 0$, $E[S_k] = k + k^2/2$, for $k = 2, \ldots, n-1$ and using that $P\{X_n = n-1\} = 0$, we see that

$$\begin{aligned} E[S_n] &= n + E[S_n]P\{X_n = 0\} + \sum_{i=1}^{n}[n - i + (n-i)^2/2]P\{X_n = i\} \\ &= n + E[S_n]P\{X_n = 0\} + (n + n^2/2)(1 - P\{X_n = 0\}) \\ &\quad -(n+1)E[X_n] + E[X_n^2]/2 \end{aligned}$$

Substituting the identities $E[X_n] = 1$, $E[X_n^2] = 2$ in the preceding shows that

$$E[S_n] = n + n^2/2$$

and the induction proof is complete.
(c) If we let C_j denote the number of hats chosen by person j, $j = 1, \ldots, n$ then

$$\sum_{j=1}^{n} C_j = S_n$$

Taking expectations, and using the fact that each C_j has the same mean, yields the result

$$E[C_j] = E[S_n]/n = 1 + n/2$$

Hence, the expected number of false selections by person j is

$$E[C_j - 1] = n/2. \qquad \blacksquare$$

Example 3.15 Independent trials, each of which is a success with probability p, are performed until there are k consecutive successes. What is the mean number of necessary trials?

Solution: Let N_k denote the number of necessary trials to obtain k consecutive successes, and let $M_k = E[N_k]$. We will determine M_k by deriving and then solving a recursive equation that it satisfies. To begin, write

$$N_k = N_{k-1} + A_{k-1,k}$$

where N_{k-1} is the number of trials needed for $k-1$ consecutive successes, and $A_{k-1,k}$ is the number of additional trials needed to go from having $k-1$ successes in a row to having k in a row. Taking expectations gives that,

$$M_k = M_{k-1} + E[A_{k-1,k}]$$

To determine $E[A_{k-1,k}]$, condition on the next trial after there have been $k-1$ successes in a row. If it is a success then that gives k in a row and no additional trials after that are needed; if it is a failure then at that point we are starting all over and so the expected addtional number from then on would be $E[N_k]$. Thus,

$$E[A_{k-1,k}] = 1 \cdot p + (1 + M_k)(1 - p) = 1 + (1 - p)M_k$$

giving that

$$M_k = M_{k-1} + 1 + (1 - p)M_k$$

or

$$M_k = \frac{1}{p} + \frac{M_{k-1}}{p}$$

Since N_1, the time of the first success, is geometric with parameter p, we see that

$$M_1 = \frac{1}{p}$$

and, recursively

$$M_2 = \frac{1}{p} + \frac{1}{p^2},$$
$$M_3 = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3}$$

and, in general,

$$M_k = \frac{1}{p} + \frac{1}{p^2} + \cdots + \frac{1}{p^k}$$ ■

Example 3.16 (Analyzing the Quick-Sort Algorithm) Suppose we are given a set of n distinct values—$x_1, \ldots, x_n$—and we desire to put these values in increasing order or, as it is commonly called, to *sort* them. An efficient procedure for accomplishing this is the quick-sort algorithm, which is defined recursively as follows: When $n = 2$ the algorithm compares the two values and puts them in the appropriate order. When $n > 2$ it starts by choosing at random one of the n values—say, x_i—and then compares each of the other $n - 1$ values with x_i, noting which are smaller and which are larger than x_i. Letting S_i denote the set of elements smaller than x_i, and $\bar{S}_i$ the set of elements greater than x_i, the algorithm now sorts the set S_i and the set $\bar{S}_i$. The final ordering, therefore, consists of the ordered set of the elements in S_i, then x_i, and then the ordered set of the elements in $\bar{S}_i$. For instance, suppose that the set of elements is 10, 5, 8, 2, 1, 4, 7. We start by choosing one of these values at random (that is, each of the 7 values has probability of $\frac{1}{7}$ of being chosen). Suppose, for instance, that the value 4 is chosen. We then compare 4 with each of the other six values to obtain

$$\{2, 1\}, 4, \{10, 5, 8, 7\}$$

We now sort the set {2, 1} to obtain

$$1, 2, 4, \{10, 5, 8, 7\}$$

Next we choose a value at random from {10, 5, 8, 7}—say 7 is chosen—and compare each of the other three values with 7 to obtain

$$1, 2, 4, 5, 7, \{10, 8\}$$

Finally, we sort {10, 8} to end up with

$$1, 2, 4, 5, 7, 8, 10$$

One measure of the effectiveness of this algorithm is the expected number of comparisons that it makes. Let us denote by M_n the expected number of comparisons needed by the quick-sort algorithm to sort a set of n distinct values. To obtain a recursion for M_n we condition on the rank of the initial value selected to obtain

$$M_n = \sum_{j=1}^{n} E[\text{number of comparisons}|\text{value selected is } j\text{th smallest}]\frac{1}{n}$$

Now, if the initial value selected is the jth smallest, then the set of values smaller than it is of size $j-1$, and the set of values greater than it is of size $n-j$. Hence, as $n-1$ comparisons with the initial value chosen must be made, we see that

$$\begin{aligned} M_n &= \sum_{j=1}^{n} (n-1+M_{j-1}+M_{n-j})\frac{1}{n} \\ &= n-1+\frac{2}{n}\sum_{k=1}^{n-1} M_k \quad (\text{since } M_0 = 0) \end{aligned}$$

or, equivalently,

$$nM_n = n(n-1) + 2\sum_{k=1}^{n-1} M_k$$

To solve the preceding, note that upon replacing n by $n+1$ we obtain

$$(n+1)M_{n+1} = (n+1)n + 2\sum_{k=1}^{n} M_k$$

Hence, upon subtraction,

$$(n+1)M_{n+1} - nM_n = 2n + 2M_n$$

or

$$(n+1)M_{n+1} = (n+2)M_n + 2n$$

Therefore,

$$\frac{M_{n+1}}{n+2} = \frac{2n}{(n+1)(n+2)} + \frac{M_n}{n+1}$$

Iterating this gives

$$\begin{aligned}\frac{M_{n+1}}{n+2} &= \frac{2n}{(n+1)(n+2)} + \frac{2(n-1)}{n(n+1)} + \frac{M_{n-1}}{n} \\ &= \cdots \\ &= 2\sum_{k=0}^{n-1} \frac{n-k}{(n+1-k)(n+2-k)} \quad \text{since } M_1 = 0\end{aligned}$$

Hence,

$$\begin{aligned}M_{n+1} &= 2(n+2)\sum_{k=0}^{n-1} \frac{n-k}{(n+1-k)(n+2-k)} \\ &= 2(n+2)\sum_{i=1}^{n} \frac{i}{(i+1)(i+2)}, \quad n \geqslant 1\end{aligned}$$

Using the identity $i/(i+1)(i+2) = 2/(i+2) - 1/(i+1)$, we can approximate M_{n+1} for large n as follows:

$$\begin{aligned}M_{n+1} &= 2(n+2)\left[\sum_{i=1}^{n} \frac{2}{i+2} - \sum_{i=1}^{n} \frac{1}{i+1}\right] \\ &\sim 2(n+2)\left[\int_3^{n+2} \frac{2}{x}\,dx - \int_2^{n+1} \frac{1}{x}\,dx\right] \\ &= 2(n+2)[2\log(n+2) - \log(n+1) + \log 2 - 2\log 3] \\ &= 2(n+2)\left[\log(n+2) + \log\frac{n+2}{n+1} + \log 2 - 2\log 3\right] \\ &\sim 2(n+2)\log(n+2)\end{aligned}$$

■

Although we usually employ the conditional expectation identity to more easily enable us to compute an unconditional expectation, in our next example we show how it can sometimes be used to obtain the conditional expectation.

Example 3.17 In the match problem of Example 2.31 involving n, $n > 1$, individuals, find the conditional expected number of matches given that the first person did not have a match.

Solution: Let X denote the number of matches, and let X_1 equal 1 if the first person has a match and 0 otherwise. Then,

$$\begin{aligned}E[X] &= E[X|X_1 = 0]P\{X_1 = 0\} + E[X|X_1 = 1]P\{X_1 = 1\} \\ &= E[X|X_1 = 0]\frac{n-1}{n} + E[X|X_1 = 1]\frac{1}{n}\end{aligned}$$

But, from Example 2.31

$$E[X] = 1$$

Moreover, given that the first person has a match, the expected number of matches is equal to 1 plus the expected number of matches when $n - 1$ people select among their own $n - 1$ hats, showing that

$$E[X|X_1 = 1] = 2$$

Therefore, we obtain the result

$$E[X|X_1 = 0] = \frac{n-2}{n-1}$$ ■

3.4.1 Computing Variances by Conditioning

Conditional expectations can also be used to compute the variance of a random variable. Specifically, we can use

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

and then use conditioning to obtain both $E[X]$ and $E[X^2]$. We illustrate this technique by determining the variance of a geometric random variable.

Example 3.18 (Variance of the Geometric Random Variable) Independent trials, each resulting in a success with probability p, are performed in sequence. Let N be the trial number of the first success. Find $\text{Var}(N)$.

Solution: Let $Y = 1$ if the first trial results in a success, and $Y = 0$ otherwise.

$$\text{Var}(N) = E[N^2] - (E[N])^2$$

To calculate $E[N^2]$ and $E[N]$ we condition on Y. For instance,

$$E[N^2] = E\big[E[N^2|Y]\big]$$

However,

$$E[N^2|Y = 1] = 1,$$
$$E[N^2|Y = 0] = E[(1 + N)^2]$$

These two equations are true since if the first trial results in a success, then clearly $N = 1$ and so $N^2 = 1$. On the other hand, if the first trial results in a failure, then the total number of trials necessary for the first success will equal one (the first trial that results in failure) plus the necessary number of additional trials. Since this latter quantity has the same distribution as N, we get that $E[N^2|Y = 0] = E[(1 + N)^2]$. Hence, we see that

$$\begin{aligned} E[N^2] &= E[N^2|Y = 1]P\{Y = 1\} + E[N^2|Y = 0]P\{Y = 0\} \\ &= p + E[(1 + N)^2](1 - p) \\ &= 1 + (1 - p)E[2N + N^2] \end{aligned}$$

Since, as was shown in Example 3.10, $E[N] = 1/p$, this yields

$$E[N^2] = 1 + \frac{2(1-p)}{p} + (1-p)E[N^2]$$

or

$$E[N^2] = \frac{2-p}{p^2}$$

Therefore,

$$\begin{aligned}\mathrm{Var}(N) &= E[N^2] - (E[N])^2 \\ &= \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 \\ &= \frac{1-p}{p^2}\end{aligned}$$

■

Another way to use conditioning to obtain the variance of a random variable is to apply the conditional variance formula. The conditional variance of X given that $Y = y$ is defined by

$$\mathrm{Var}(X|Y=y) = E\big[(X - E[X|Y=y])^2|Y=y\big]$$

That is, the conditional variance is defined in exactly the same manner as the ordinary variance with the exception that all probabilities are determined conditional on the event that $Y = y$. Expanding the right side of the preceding and taking expectation term by term yields

$$\mathrm{Var}(X|Y=y) = E[X^2|Y=y] - (E[X|Y=y])^2$$

Letting $\mathrm{Var}(X|Y)$ denote that function of Y whose value when $Y = y$ is $\mathrm{Var}(X|Y=y)$, we have the following result.

Proposition 3.1 (The Conditional Variance Formula)

$$\mathrm{Var}(X) = E\big[\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big(E[X|Y]\big) \tag{3.7}$$

Proof.

$$\begin{aligned}E\big[\mathrm{Var}(X|Y)\big] &= E[E[X^2|Y] - (E[X|Y])^2] \\ &= E\big[E[X^2|Y]\big] - E\big[(E[X|Y])^2\big] \\ &= E[X^2] - E\big[(E[X|Y])^2\big]\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(E[X|Y]) &= E\big[(E[X|Y])^2\big] - \big(E\big[E[X|Y]\big]\big)^2 \\ &= E\big[(E[X|Y])^2\big] - (E[X])^2\end{aligned}$$

Therefore,

$$E\big[\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big(E[X|Y]\big) = E[X^2] - (E[X])^2$$

which completes the proof. ■

Example 3.19 (The Variance of a Compound Random Variable) Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with distribution F having mean μ and variance σ^2, and assume that they are independent of the nonnegative integer valued random variable N. As noted in Example 3.10, where its expected value was determined, the random variable $S = \sum_{i=1}^{N} X_i$ is called a compound random variable. Find its variance.

Solution: Whereas we could obtain $E[S^2]$ by conditioning on N, let us instead use the conditional variance formula. Now,

$$\begin{aligned}
\mathrm{Var}(S|N = n) &= \mathrm{Var}\left(\sum_{i=1}^{N} X_i | N = n\right) \\
&= \mathrm{Var}\left(\sum_{i=1}^{n} X_i | N = n\right) \\
&= \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= n\sigma^2
\end{aligned}$$

By the same reasoning,

$$E[S|N = n] = n\mu$$

Therefore,

$$\mathrm{Var}(S|N) = N\sigma^2, \quad E[S|N] = N\mu$$

and the conditional variance formula gives

$$\mathrm{Var}(S) = E[N\sigma^2] + \mathrm{Var}(N\mu) = \sigma^2 E[N] + \mu^2 \mathrm{Var}(N)$$

If N is a Poisson random variable, then $S = \sum_{i=1}^{N} X_i$ is called a *compound Poisson* random variable. Because the variance of a Poisson random variable is equal to its mean, it follows that for a compound Poisson random variable having $E[N] = \lambda$

$$\mathrm{Var}(S) = \lambda\sigma^2 + \lambda\mu^2 = \lambda E[X^2]$$

where X has the distribution F. ■

Example 3.20 (The Variance in the Matching Rounds Problem) Consider the matching rounds problem of Example 3.14, and let $V_n = \mathrm{Var}(R_n)$ denote the variance

of the number of rounds needed when there are initially n people. Using the conditional variance formula, we will show that

$$V_n = n, \quad n \geqslant 2$$

The proof of the preceding is by induction on n. To begin, note that when $n = 2$ the number of rounds needed is geometric with parameter $p = 1/2$ and so

$$V_2 = \frac{1-p}{p^2} = 2$$

So assume the induction hypothesis that

$$V_j = j, \quad 2 \leqslant j < n$$

and now consider the case when there are n individuals. If X is the number of matches in the first round then, conditional on X, the number of rounds R_n is distributed as 1 plus the number of rounds needed when there are initially $n - X$ individuals. Consequently,

$$\begin{aligned} E[R_n|X] &= 1 + E[R_{n-X}] \\ &= 1 + n - X \qquad \text{by Example 3.14} \end{aligned}$$

Also, with $V_0 = 0$,

$$\text{Var}(R_n|X) = \text{Var}(R_{n-X}) = V_{n-X}$$

Hence, by the conditional variance formula

$$\begin{aligned} V_n &= E[\text{Var}(R_n|X)] + \text{Var}(E[R_n|X]) \\ &= E[V_{n-X}] + \text{Var}(X) \\ &= \sum_{j=0}^{n} V_{n-j} P(X = j) + \text{Var}(X) \\ &= V_n P(X = 0) + \sum_{j=1}^{n} V_{n-j} P(X = j) + \text{Var}(X) \end{aligned}$$

Because $P(X = n-1) = 0$, it follows from the preceding and the induction hypothesis that

$$\begin{aligned} V_n &= V_n P(X = 0) + \sum_{j=1}^{n} (n-j) P(X = j) + \text{Var}(X) \\ &= V_n P(X = 0) + n(1 - P(X = 0)) - E[X] + \text{Var}(X) \end{aligned}$$

As it is easily shown (see Example 2.31 and Exercise 72 of Chapter 2) that $E[X] = \text{Var}(X) = 1$, the preceding gives

$$V_n = V_n P(X = 0) + n(1 - P(X = 0))$$

thus proving the result. ■