6.3 SUMS OF INDEPENDENT RANDOM VARIABLES

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 that X and Y are independent, continuous random variables having probability density functions f_X and f_Y. The cumulative distribution function of $X + Y$ is obtained as follows:

$$\begin{aligned} F_{X+Y}(a) &= P\{X + Y \le a\} \\ &= \iint\limits_{x+y\le a} f_X(x) f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x) f_Y(y)\,dx\,dy \qquad (3.1) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)\,dx f_Y(y)\,dy \\ &= \int_{-\infty}^{\infty} F_X(a - y) f_Y(y)\,dy \end{aligned}$$

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 (3.1), we find that the probability density function f_{X+Y} of $X + Y$ is given by

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da}\int_{-\infty}^{\infty} F_X(a - y) f_Y(y)\,dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da} F_X(a - y) f_Y(y)\,dy \qquad (3.2) \\ &= \int_{-\infty}^{\infty} f_X(a - y) f_Y(y)\,dy \end{aligned}$$

6.3.1 Identically Distributed Uniform Random Variables

It is not difficult to determine the density function of the sum of two independent uniform $(0, 1)$ random variables.

EXAMPLE 3a Sum of two independent uniform random variables

If X and Y are independent random variables, both uniformly distributed on (0, 1), calculate the probability density of $X + Y$.

Solution. From Equation (3.2), since

$$f_X(a) = f_Y(a) = \begin{cases} 1 & 0 < a < 1 \\ 0 & \text{otherwise} \end{cases}$$

we obtain

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

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

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

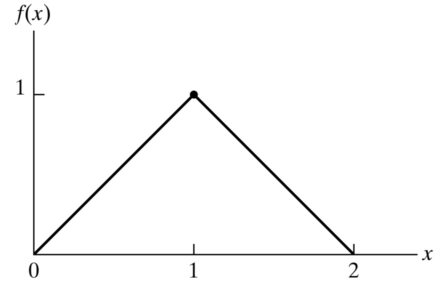

FIGURE 6.3: Triangular density function.

For $1 < a < 2$, we get

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

Hence,

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

Because of the shape of its density function (see Figure 6.3), the random variable $X + Y$ is said to have a *triangular* distribution. ■

Now, suppose that $X_1, X_2, \ldots, X_n$ are independent uniform $(0, 1)$ random variables, and let

$$F_n(x) = P\{X_1 + \ldots + X_n \le x\}$$

Whereas a general formula for $F_n(x)$ is messy, it has a particularly nice form when $x \le 1$. Indeed, we now use mathematical induction to prove that

$$F_n(x) = x^n/n!\,, \quad 0 \le x \le 1$$

Because the proceeding equation is true for $n = 1$, assume that

$$F_{n-1}(x) = x^{n-1}/(n - 1)!\,, \quad 0 \le x \le 1$$

Now, writing

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n$$

and using the fact that the X_i are all nonnegative, we see from Equation 3.1 that, for $0 \le x \le 1$,

$$\begin{aligned} F_n(x) &= \int_0^1 F_{n-1}(x - y) f_{X_n}(y)dy \\ &= \frac{1}{(n - 1)!} \int_0^x (x - y)^{n-1}\, dy \quad \text{by the induction hypothesis} \\ &= x^n/n! \end{aligned}$$

which completes the proof.

For an interesting application of the preceding formula, let us use it to determine the expected number of independent uniform (0, 1) random variables that need to be summed to exceed 1. That is, with $X_1, X_2, \ldots$ being independent uniform (0, 1) random variables, we want to determine $E[N]$, where

$$N = \min\{n : X_1 + \ldots + X_n > 1\}$$

Noting that N is greater than $n > 0$ if and only if $X_1 + \ldots + X_n \leq 1$, we see that

$$P\{N > n\} = F_n(1) = 1/n!\,, \quad n > 0$$

Because

$$P\{N > 0\} = 1 = 1/0!$$

we see that, for $n > 0$,

$$P\{N = n\} = P\{N > n - 1\} - P\{N > n\} = \frac{1}{(n-1)!} - \frac{1}{n!} = \frac{n-1}{n!}$$

Therefore,

$$\begin{aligned} E[N] &= \sum_{n=1}^{\infty} \frac{n(n-1)}{n!} \\ &= \sum_{n=2}^{\infty} \frac{1}{(n-2)!} \\ &= e \end{aligned}$$

That is, the mean number of independent uniform (0, 1) random variables that must be summed for the sum to exceed 1 is equal to e.

6.3.2 Gamma Random Variables

Recall that a gamma random variable has a density of the form

$$f(y) = \frac{\lambda e^{-\lambda y}(\lambda y)^{t-1}}{\Gamma(t)} \quad 0 < y < \infty$$

An important property of this family of distributions is that, for a fixed value of λ, it is closed under convolutions.

Proposition 3.1. If X and Y are independent gamma random variables with respective parameters (s, λ) and (t, λ), then $X + Y$ is a gamma random variable with parameters $(s + t, \lambda)$.

Proof. Using Equation (3.2), we obtain

$$\begin{aligned} f_{X+Y}(a) &= \frac{1}{\Gamma(s)\Gamma(t)} \int_0^a \lambda e^{-\lambda(a-y)}[\lambda(a - y)]^{s-1} \lambda e^{-\lambda y}(\lambda y)^{t-1}\,dy \\ &= Ke^{-\lambda a} \int_0^a (a - y)^{s-1} y^{t-1} dy \\ &= Ke^{-\lambda a} a^{s+t-1} \int_0^1 (1 - x)^{s-1} x^{t-1}\,dx \quad \text{by letting } x = \frac{y}{a} \\ &= Ce^{-\lambda a} a^{s+t-1} \end{aligned}$$

where C is a constant that does not depend on a. But, as the preceding is a density function and thus must integrate to 1, the value of C is determined, and we have

$$f_{X+Y}(a) = \frac{\lambda e^{-\lambda a}(\lambda a)^{s+t-1}}{\Gamma(s + t)}$$

Hence, the result is proved. □

It is now a simple matter to establish, by using Proposition 3.1 and induction, that if $X_i, i = 1, \ldots, n$ are independent gamma random variables with respective parameters $(t_i, \lambda), i = 1, \ldots, n$, then $\sum_{i=1}^{n} X_i$ is gamma with parameters $\left(\sum_{i=1}^{n} t_i, \lambda\right)$. We leave the proof of this statement as an exercise.

EXAMPLE 3b

Let $X_1, X_2, \ldots, X_n$ be n independent exponential random variables, each having parameter λ. Then, since an exponential random variable with parameter λ is the same as a gamma random variable with parameters $(1, \lambda)$, it follows from Proposition 3.1 that $X_1 + X_2 + \cdots + X_n$ is a gamma random variable with parameters (n, λ). ■

If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables, then $Y \equiv \sum_{i=1}^{n} Z_i^2$ is said to have the *chi-squared* (sometimes seen as χ^2) distribution with n degrees of freedom. Let us compute the density function of Y. When $n = 1, Y = Z_1^2$, and from Example 7b of Chapter 5, we see that its probability density function is given by

$$\begin{aligned} f_{Z^2}(y) &= \frac{1}{2\sqrt{y}}[f_Z(\sqrt{y}) + f_Z(-\sqrt{y})] \\ &= \frac{1}{2\sqrt{y}}\frac{2}{\sqrt{2\pi}}e^{-y/2} \\ &= \frac{\frac{1}{2}e^{-y/2}(y/2)^{1/2-1}}{\sqrt{\pi}} \end{aligned}$$

But we recognize the preceding as the gamma distribution with parameters $\left(\frac{1}{2}, \frac{1}{2}\right)$. [A by-product of this analysis is that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.] But since each Z_i^2 is gamma $\left(\frac{1}{2}, \frac{1}{2}\right)$, it follows from Proposition 3.1 that the χ^2 distribution with n degrees of freedom is just the gamma distribution with parameters $\left(n/2, \frac{1}{2}\right)$ and hence has a probability density function given by

$$\begin{aligned} f_{\chi^2}(y) &= \frac{\frac{1}{2}e^{-y/2}\left(\frac{y}{2}\right)^{n/2-1}}{\Gamma\left(\frac{n}{2}\right)} \qquad y > 0 \\ &= \frac{e^{-y/2}y^{n/2-1}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \qquad y > 0 \end{aligned}$$

When n is an even integer, $\Gamma(n/2) = [(n/2) - 1]!$, whereas when n is odd, $\Gamma(n/2)$ can be obtained from iterating the relationship $\Gamma(t) = (t - 1)\Gamma(t - 1)$ and then using the previously obtained result that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. [For instance, $\Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3}{2}\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{3}{4}\sqrt{\pi}$.]

In practice, the chi-squared distribution often arises as the distribution of the square of the error involved when one attempts to hit a target in n-dimensional space when the coordinate errors are taken to be independent standard normal random variables. It is also important in statistical analysis.

6.3.3 Normal Random Variables

We can also use Equation (3.2) to prove the following important result about normal random variables.

Proposition 3.2. If $X_i, i = 1, \ldots, n$, are independent random variables that are normally distributed with respective parameters $\mu_i, \sigma_i^2, i = 1, \ldots, n$, then $\sum_{i=1}^{n} X_i$ is normally distributed with parameters $\sum_{i=1}^{n} \mu_i$ and $\sum_{i=1}^{n} \sigma_i^2$.

Proof of Proposition 3.2: To begin, let X and Y be independent normal random variables with X having mean 0 and variance σ^2 and Y having mean 0 and variance 1. We will determine the density function of $X + Y$ by utilizing Equation (3.2). Now, with

$$c = \frac{1}{2\sigma^2} + \frac{1}{2} = \frac{1 + \sigma^2}{2\sigma^2}$$

we have

$$\begin{aligned}
f_X(a - y)f_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(a - y)^2}{2\sigma^2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \\
&= \frac{1}{2\pi\sigma} \exp\left\{-\frac{a^2}{2\sigma^2}\right\} \exp\left\{-c\left(y^2 - 2y\frac{a}{1 + \sigma^2}\right)\right\}
\end{aligned}$$

Hence, from Equation (3.2),

$$\begin{aligned}
f_{X+Y}(a) &= \frac{1}{2\pi\sigma} \exp\left\{-\frac{a^2}{2\sigma^2}\right\} \exp\left\{\frac{a^2}{2\sigma^2(1 + \sigma^2)}\right\} \\
&\quad \times \int_{-\infty}^{\infty} \exp\left\{-c\left(y - \frac{a}{1 + \sigma^2}\right)^2\right\} dy \\
&= \frac{1}{2\pi\sigma} \exp\left\{-\frac{a^2}{2(1 + \sigma^2)}\right\} \int_{-\infty}^{\infty} \exp\{-cx^2\}\, dx \\
&= C \exp\left\{-\frac{a^2}{2(1 + \sigma^2)}\right\}
\end{aligned}$$

where C does not depend on a. But this implies that $X + Y$ is normal with mean 0 and variance $1 + \sigma^2$.

Now, suppose that X_1 and X_2 are independent normal random variables with X_i having mean μ_i and variance $\sigma_i^2, i = 1, 2$. Then

$$X_1 + X_2 = \sigma_2 \left(\frac{X_1 - \mu_1}{\sigma_2} + \frac{X_2 - \mu_2}{\sigma_2} \right) + \mu_1 + \mu_2$$

But since $(X_1 - \mu_1)/\sigma_2$ is normal with mean 0 and variance σ_1^2/σ_2^2, and $(X_2 - \mu_2)/\sigma_2$ is normal with mean 0 and variance 1, it follows from our previous result that $(X_1 - \mu_1)/\sigma_2 + (X_2 - \mu_2)/\sigma_2$ is normal with mean 0 and variance $1 + \sigma_1^2/\sigma_2^2$, implying that $X_1 + X_2$ is normal with mean $\mu_1 + \mu_2$ and variance $\sigma_2^2(1 + \sigma_1^2/\sigma_2^2) = \sigma_1^2 + \sigma_2^2$.

Thus, Proposition 3.2 is established when $n = 2$. The general case now follows by induction. That is, assume that Proposition 3.2 is true when there are $n - 1$ random variables. Now consider the case of n, and write

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n$$

By the induction hypothesis, $\sum_{i=1}^{n-1} X_i$ is normal with mean $\sum_{i=1}^{n-1} \mu_i$ and variance $\sum_{i=1}^{n-1} \sigma_i^2$. Therefore, by the result for $n = 2$, $\sum_{i=1}^{n} X_i$ is normal with mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$.

EXAMPLE 3c

A basketball team will play a 44-game season. Twenty-six of these games are against class A teams and 18 are against class B teams. Suppose that the team will win each game against a class A team with probability .4 and will win each game against a class B team with probability .7. Suppose also that the results of the different games are independent. Approximate the probability that

(a) the team wins 25 games or more;
(b) the team wins more games against class A teams than it does against class B teams.

Solution. (a) Let X_A and X_B respectively denote the number of games the team wins against class A and against class B teams. Note that X_A and X_B are independent binomial random variables and

$$E[X_A] = 26(.4) = 10.4 \quad \text{Var}(X_A) = 26(.4)(.6) = 6.24$$
$$E[X_B] = 18(.7) = 12.6 \quad \text{Var}(X_B) = 18(.7)(.3) = 3.78$$

By the normal approximation to the binomial, X_A and X_B will have approximately the same distribution as would independent normal random variables with the preceding expected values and variances. Hence, by Proposition 3.2, $X_A + X_B$ will have approximately a normal distribution with mean 23 and variance 10.02. Therefore, letting Z denote a standard normal random variable, we have

$$\begin{aligned} P\{X_A + X_B \geq 25\} &= P\{X_A + X_B \geq 24.5\} \\ &= P\left\{\frac{X_A + X_B - 23}{\sqrt{10.02}} \geq \frac{24.5 - 23}{\sqrt{10.02}}\right\} \\ &\approx P\left\{Z \geq \frac{1.5}{\sqrt{10.02}}\right\} \\ &\approx 1 - P\{Z < .4739\} \\ &\approx .3178 \end{aligned}$$

(b) We note that $X_A - X_B$ will have approximately a normal distribution with mean -2.2 and variance 10.02. Hence,

$$\begin{aligned} P\{X_A - X_B \geq 1\} &= P\{X_A - X_B \geq .5\} \\ &= P\left\{\frac{X_A - X_B + 2.2}{\sqrt{10.02}} \geq \frac{.5 + 2.2}{\sqrt{10.02}}\right\} \\ &\approx P\left\{Z \geq \frac{2.7}{\sqrt{10.02}}\right\} \\ &\approx 1 - P\{Z < .8530\} \\ &\approx .1968 \end{aligned}$$

Therefore, there is approximately a 31.78 percent chance that the team will win at least 25 games and approximately a 19.68 percent chance that it will win more games against class A teams than against class B teams. ■

The random variable Y is said to be a *lognormal* random variable with parameters μ and σ if log (Y) is a normal random variable with mean μ and variance σ^2. That is, Y is lognormal if it can be expressed as

$$Y = e^X$$

where X is a normal random variable.

EXAMPLE 3d

Starting at some fixed time, let $S(n)$ denote the price of a certain security at the end of n additional weeks, $n \geq 1$. A popular model for the evolution of these prices assumes that the price ratios $S(n)/S(n-1), n \geq 1$, are independent and identically distributed lognormal random variables. Assuming this model, with parameters $\mu = .0165, \sigma = .0730$, what is the probability that

(a) the price of the security increases over each of the next two weeks?
(b) the price at the end of two weeks is higher than it is today?

Solution. Let Z be a standard normal random variable. To solve part (a), we use the fact that $\log(x)$ increases in x to conclude that $x > 1$ if and only if $\log(x) > \log(1) = 0$. As a result, we have

$$P\left\{\frac{S(1)}{S(0)} > 1\right\} = P\left\{\log\left(\frac{S(1)}{S(0)}\right) > 0\right\}$$
$$= P\left\{Z > \frac{-.0165}{.0730}\right\}$$
$$= P\{Z < .2260\}$$
$$= .5894$$

In other words, the probability that the price is up after one week is .5894. Since the successive price ratios are independent, the probability that the price increases over each of the next two weeks is $(.5894)^2 = .3474$.

To solve part (b), we reason as follows:

$$P\left\{\frac{S(2)}{S(0)} > 1\right\} = P\left\{\frac{S(2)}{S(1)}\frac{S(1)}{S(0)} > 1\right\}$$
$$= P\left\{\log\left(\frac{S(2)}{S(1)}\right) + \log\left(\frac{S(1)}{S(0)}\right) > 0\right\}$$

However, $\log\left(\frac{S(2)}{S(1)}\right) + \log\left(\frac{S(1)}{S(0)}\right)$, being the sum of two independent normal random variables with a common mean .0165 and a common standard deviation .0730, is a normal random variable with mean .0330 and variance $2(.0730)^2$. Consequently,

$$P\left\{\frac{S(2)}{S(0)} > 1\right\} = P\left\{Z > \frac{-.0330}{.0730\sqrt{2}}\right\}$$
$$= P\{Z < .31965\}$$
$$= .6254$$

■

6.3.4 Poisson and Binomial Random Variables

Rather than attempt to derive a general expression for the distribution of $X + Y$ in the discrete case, we shall consider some examples.

EXAMPLE 3e Sums of independent Poisson random variables

If X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2, compute the distribution of $X + Y$.

Solution. Because 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

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

$$= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!}$$
$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}$$
$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n$$

Thus, $X_1 + X_2$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$. ■

EXAMPLE 3f Sums of independent binomial random variables

Let X and Y be independent binomial random variables with respective parameters (n, p) and (m, p). Calculate the distribution of $X + Y$.

Solution. Recalling the interpretation of a binomial random variable, and without any computation at all, we can immediately conclude that $X + Y$ is binomial with parameters $(n + m, p)$. This follows because X represents the number of successes in n independent trials, each of which results in a success with probability p; similarly, Y represents the number of successes in m independent trials, each of which results in a success with probability p. Hence, given that X and Y are assumed independent, it follows that $X + Y$ represents the number of successes in $n + m$ independent trials when each trial has a probability p of resulting in a success. Therefore, $X + Y$ is a binomial random variable with parameters $(n + m, p)$. To check this conclusion analytically, note that

$$P\{X + Y = k\} = \sum_{i=0}^{n} P\{X = i, Y = k - i\}$$
$$= \sum_{i=0}^{n} P\{X = i\}P\{Y = k - i\}$$
$$= \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \binom{m}{k-i} p^{k-i} q^{m-k+i}$$

where $q = 1 - p$ and where $\binom{r}{j} = 0$ when $j < 0$. Thus,

$$P\{X + Y = k\} = p^k q^{n+m-k} \sum_{i=0}^{n} \binom{n}{i} \binom{m}{k-i}$$

and the conclusion follows upon application of the combinatorial identity

$$\binom{n+m}{k} = \sum_{i=0}^{n} \binom{n}{i} \binom{m}{k-i}$$

■

6.3.5 Geometric Random Variables

Let $X_1, \ldots, X_n$ be independent geometric random variables, with X_i having parameter p_i for $i = 1, \ldots, n$. We are interested in computing the probability mass function of their sum $S_n = \sum_{i=1}^{n} X_i$. For an application, consider n coins, with coin i having probability p_i of coming up heads when flipped, $i = 1, \ldots, n$. Suppose that coin 1 is flipped until heads appears, at which point coin 2 is flipped until it shows heads, and then coin 3 is flipped until it shows heads, and so on. If we let X_i denote the number of flips made with coin i, then $X_1, X_2, \ldots, X_n$ will be independent geometric random variables with respective parameters $p_1, p_2, \ldots, p_n$, and $S_n = \sum_{i=1}^{n} X_i$ will represent the total number of flips. If all the p_i are equal—say, all $p_i = p$—then S_n has the same distribution as the number of flips of a coin having probability p of coming up heads that are needed to obtain a total of n heads, and so S_n is a negative binomial random variable with probability mass function

$$P\{S_n = k\} = \binom{k-1}{n-1} p^n (1-p)^{k-n}, \quad k \geq n$$

As a prelude to determining the probability mass function of S_n when the p_i are all distinct, let us first consider the case $n = 2$. Letting $q_j = 1 - p_j$, $j = 1, 2$, we obtain

$$\begin{aligned}
P(S_2 = k) &= \sum_{j=1}^{k-1} P\{X_1 = j, X_2 = k - j\} \\
&= \sum_{j=1}^{k-1} P\{X_1 = j\}\, P\{X_2 = k - j\} \quad \text{(by independence)} \\
&= \sum_{j=1}^{k-1} p_1 q_1^{j-1} p_2 q_2^{k-j-1} \\
&= p_1 p_2 q_2^{k-2} \sum_{j=1}^{k-1} (q_1/q_2)^{j-1} \\
&= p_1 p_2 q_2^{k-2} \frac{1 - (q_1/q_2)^{k-1}}{1 - q_1/q_2} \\
&= \frac{p_1 p_2 q_2^{k-1}}{q_2 - q_1} - \frac{p_1 p_2 q_1^{k-1}}{q_2 - q_1} \\
&= p_2 q_2^{k-1} \frac{p_1}{p_1 - p_2} + p_1 q_1^{k-1} \frac{p_2}{p_2 - p_1}
\end{aligned}$$

If we now let $n = 3$ and compute $P\{S_3 = k\}$ by starting with the identity

$$P\{S_3 = k\} = \sum_{j=1}^{k-1} P\{S_2 = j, X_3 = k - j\} = \sum_{j=1}^{k-1} P\{S_2 = j\} P\{X_3 = k - j\}$$

and then substituting the derived formula for the mass function of S_2, we would obtain, after some computations,

$$\begin{aligned}
P\{S_3 = k\} = {} & p_1 q_1^{k-1} \frac{p_2}{p_2 - p_1} \frac{p_3}{p_3 - p_1} + p_2 q_2^{k-1} \frac{p_1}{p_1 - p_2} \frac{p_3}{p_3 - p_2} \\
& + p_3 q_3^{k-1} \frac{p_1}{p_1 - p_3} \frac{p_2}{p_2 - p_3}
\end{aligned}$$

The mass functions of S_2 and S_3 lead to the following conjecture for the mass function of S_n.

Proposition 3.3. Let $X_1, \ldots, X_n$ be independent geometric random variables, with X_i having parameter p_i for $i = 1, \ldots, n$. If all the p_i are distinct, then, for $k \geq n$,

$$P\{S_n = k\} = \sum_{i=1}^{n} p_i q_i^{k-1} \prod_{j \neq i} \frac{p_j}{p_j - p_i}$$

Proof of Proposition 3.3: We will prove this proposition by induction on the value of $n + k$. Because the proposition is true when $n = 2, k = 2$, take as the induction hypothesis that it is true for any $k \geq n$ for which $n + k \leq r$. Now, suppose $k \geq n$ are such that $n + k = r + 1$. To compute $P\{S_n = k\}$, we condition on whether $X_n = 1$. This gives

$$\begin{aligned} P\{S_n = k\} &= P\{S_n = k | X_n = 1\}P\{X_n = 1\} + P\{S_n = k | X_n > 1\}P\{X_n > 1\} \\ &= P\{S_n = k | X_n = 1\}p_n + P\{S_n = k | X_n > 1\}q_n \end{aligned}$$

Now,

$$\begin{aligned} P\{S_n = k | X_n = 1\} &= P\{S_{n-1} = k - 1 | X_n = 1\} \\ &= P\{S_{n-1} = k - 1\} \quad \text{(by independence)} \\ &= \sum_{i=1}^{n-1} p_i q_i^{k-2} \prod_{i \neq j \leq n-1} \frac{p_j}{p_j - p_i} \quad \text{(by the induction hypothesis)} \end{aligned}$$

Now, if X is geometric with parameter p, then the conditional distribution of X given that it is larger than 1 is the same as the distribution of 1 (the first failed trial) plus a geometric with parameter p (the number of additional trials after the first until a success occurs). Consequently,

$$\begin{aligned} P\{S_n = k | X_n > 1\} &= P\{X_1 + \ldots + X_{n-1} + X_n + 1 = k\} \\ &= P\{S_n = k - 1\} \\ &= \sum_{i=1}^{n} p_i q_i^{k-2} \prod_{i \neq j \leq n} \frac{p_j}{p_j - p_i} \end{aligned}$$

where the final equality follows from the induction hypothesis. Thus, from the preceding, we obtain

$$\begin{aligned} P\{S_n = k\} &= p_n \sum_{i=1}^{n-1} p_i q_i^{k-2} \prod_{i \neq j \leq n-1} \frac{p_j}{p_j - p_i} + q_n \sum_{i=1}^{n} p_i q_i^{k-2} \prod_{i \neq j \leq n} \frac{p_j}{p_j - p_i} \\ &= p_n \sum_{i=1}^{n-1} p_i q_i^{k-2} \prod_{i \neq j \leq n-1} \frac{p_j}{p_j - p_i} + q_n \sum_{i=1}^{n-1} p_i q_i^{k-2} \prod_{i \neq j \leq n} \frac{p_j}{p_j - p_i} \\ &\quad + q_n p_n q_n^{k-2} \prod_{j < n} \frac{p_j}{p_j - p_n} \\ &= \sum_{i=1}^{n-1} p_i q_i^{k-2} p_n \left(1 + \frac{q_n}{p_n - p_i}\right) \prod_{i \neq j \leq n-1} \frac{p_j}{p_j - p_i} + p_n q_n^{k-1} \prod_{j < n} \frac{p_j}{p_j - p_n} \end{aligned}$$

Now, using that

$$1 + \frac{q_n}{p_n - p_i} = \frac{p_n - p_i + q_n}{p_n - p_i} = \frac{q_i}{p_n - p_i}$$

the preceding gives

$$\begin{aligned} P\{S_n = k\} &= \sum_{i=1}^{n-1} p_i q_i^{k-1} \prod_{i \neq j \leq n} \frac{p_j}{p_j - p_i} + p_n q_n^{k-1} \prod_{j<n} \frac{p_j}{p_j - p_n} \\ &= \sum_{i=1}^{n} p_i q_i^{k-1} \prod_{j \neq i} \frac{p_j}{p_j - p_i} \end{aligned}$$

and the proof by induction is complete. ■

6.4 CONDITIONAL DISTRIBUTIONS: DISCRETE CASE

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

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

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

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

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

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

In other words, the definitions are exactly the same as in the unconditional case, except that everything is now conditional on the event that $Y = y$. If X is independent of Y, then the conditional mass function and the distribution function are the same as the respective unconditional ones. This follows because if X is independent of Y, then

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