3

DISCRETE RANDOM VARIABLES

In the previous chapter many different distributions were developed out of Bernoulli trials. In that chapter we proceeded by creating new sample spaces for each new distribution, but when faced with many questions related to the same basic framework, it is usually clearer to maintain a single sample space and to define functions on that space whose outputs relate them to questions under consideration. Such functions are known as "random variables" and they will be the focus of this chapter.

3.1 RANDOM VARIABLES AS FUNCTIONS

EXAMPLE 3.1.1. Suppose a coin is flipped three times. Consider the probabilities associated with the following two questions:

(a) How many coins will come up heads?

(b) Which will be the first flip (if any) that shows heads?

At this point the answers to these questions should be easy to determine, but the purpose of this example is to emphasize how functions could be used to answer both within the context of a single sample space. Let S be a listing of all eight possible orderings of heads and tails on the three flips.

$$S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$$

Now define two functions on S. Let X be the function that describes the total number of heads among the three flips and let Y be the function that describes the first flip that produces heads. So,

ω	$X(\omega)$	$Y(\omega)$
hhh	3	1
hht	2	1
hth	2	1
htt	1	1
thh	2	2
tht	1	2
tth	1	3
ttt	0	none

where $Y(ttt)$ is defined as "none" since there is no first time the coin produces heads.

Suppose we want to know the probability that exactly two of the three coins will be heads. The relevant event is $E = \{hht, hth, thh\}$, but in the pre-image notation of function theory this set may also be described as $X^{-1}(\{2\})$, the elements of S for which X produces an output of 2. This allows us to describe the probability of the event as:

$$P(\text{two heads}) = P(X^{-1}(\{2\})) = P(\{hht, hth, thh\}) = \frac{3}{8}$$

Rather than use the standard pre-image notation, it is more common in probability to write $(X = 2)$ for the set $X^{-1}(\{2\})$ since this emphasizes that we are considering outcomes for which the function X equals 2.

Similarly, if we wanted to know the probability that the first result of heads showed up on the third flip, that is a question that involves the function Y. Using the notation $(Y = 3)$ in place of $Y^{-1}(\{3\})$ the probability may be calculated as

$$P(\text{first heads on flip three}) = P(Y = 3) = P(\{tth\}) = \frac{1}{8}$$

As above we can compute the

$$P(X = 0) = \frac{1}{8}, P(X = 1) = \frac{3}{8}, \text{ and } P(X = 3) = \frac{1}{8}$$

and the

$$P(Y = 1) = \frac{1}{2}, P(Y = 2) = \frac{1}{4}, \text{ and } P(Y = \text{none}) = \frac{1}{8}.$$

Thus giving a complete description of how X and Y distribute the probabilities onto their range. In both cases only a single sample space was needed. Two different questions were approached by defining two different functions on that sample space. ■

The following theorem explains how the mechanism of the previous example may be more generally applied.

THEOREM 3.1.2. *Let S be a sample space with probability P and let $X : S \to T$ be a function. Then X generates a probability Q on T given by*

$$Q(B) = P(X^{-1}(B))$$

The probability Q is called the "distribution of X" since it describes how X distributes the probability from S onto T.

The proof relies on two set-theoretic facts that we will take as given. The first is that $X^{-1}(\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} X^{-1}(B_i)$ and the second is the fact that if B_i and B_j are disjoint, then so are $X^{-1}(B_i)$ and $X^{-1}(B_j)$.

Proof - Let $B \subset T$. Since P is known to be a probability, $0 \leq P(X^{-1}(B)) \leq 1$, and so Q maps subsets of T into $[0, 1]$. Since X is a function into T, we know $X^{-1}(T) = S$. Therefore $Q(T) = P(X^{-1}(T)) = P(S) = 1$ and Q satisfies the first probability axiom.

To show Q satisfies the second axiom, suppose $B_1, B_2, \ldots$ are a countable collection of disjoint subsets of T.

$$\begin{aligned} Q(\bigcup_{i=1}^{\infty} B_i) &= P(X^{-1}(\bigcup_{i=1}^{\infty} B_i)) \\ &= P(\bigcup_{i=1}^{\infty} X^{-1}(B_i)) \\ &= \sum_{i=1}^{\infty} P(X^{-1}(B_i)) \\ &= \sum_{i=1}^{\infty} Q(B_i) \end{aligned}$$

As in the previous example, it is typical to write $(X \in B)$ in place of the notation $X^{-1}(B)$ to emphasize the fact that we are computing the probability that the function X takes a value in the set B. In practice, the new probability Q would rarely be used explicity, but would be calculated in terms of the original probability P via the relationship described in the theorem.

EXAMPLE 3.1.3. A board game has a wheel that is to be spun periodically. The wheel can stop in one of ten equally likely spots. Four of these spots are red, three are blue, two are green, and one is black. Let X denote the color of the spot. Determine the distribution of X.

The function X is defined on a sample space S that consists of the ten spots the wheel could stop, and it takes values on the set of colors $T = \{red, blue, green, black\}$. Its distribution is a probability Q on the set of colors which can be determined by calculating the probability of each color.

For instance $Q(\{red\}) = P(X = red) = P(X^{-1}(\{red\})) = \frac{4}{10}$ since four of the ten spots on the wheel are red and all spots are equally likely. Similarly,

$$\begin{aligned} Q(\{blue\}) &= P(X = blue) = \frac{3}{10} \\ Q(\{green\}) &= P(X = green) = \frac{2}{10} \\ Q(\{black\}) &= P(X = black) = \frac{1}{10} \end{aligned}$$

completing the description of the distribution. ■

EXAMPLE 3.1.4. For a certain lottery, a three-digit number is randomly selected (from 000 to 999). If a ticket matches the number exactly, it is worth \$200. If the ticket matches exactly two of the three digits, it is worth \$20. Otherwise it is worth nothing. Let X be the value of the ticket. Find the distribution of X.

The function X is defined on $S = \{000, 001, \ldots, 998, 999\}$ - the set of all one thousand possible three digit numbers. The function takes values on the set $\{0, 20, 200\}$, so the distribution Q is a probability on $T = \{0, 20, 200\}$.

First, $Q(\{200\}) = P(X = 200) = \frac{1}{1000}$ since only one of the one thousand three digit numbers is going to be an exact match.

Next, $Q(\{20\}) = P(X = 20)$, so it must be determined how many of the one thousand possibilities will have exactly two matches. There are $\binom{3}{2} = 3$ different ways to choose the two digits that will match. Those digits are determined at that point and the remaining digit must be one of the nine digits that do not match the third spot, so there are $3 \cdot 9 = 27$ three digit numbers that match exactly two digits. So $Q(\{20\}) = P(X = 20) = \frac{27}{1000}$.

Finally, since Q is a probability, $Q(\{0\}) = 1 - Q(\{20\}) - Q(\{200\}) = \frac{972}{1000}$. ■

It is frequently the case that we are interested in functions that have real-valued outputs and we reserve the term "random variable" for such a situation.

DEFINITION 3.1.5. *A "discrete random variable" is a function $X : S \to T$ where S is a sample space equipped with a probability P, and T is a countable (or finite) subset of the real numbers.*

From Theorem 3.1.2, P generates a probability on T and since it is a discrete space, the distribution may be determined by knowing the likelihood of each possible value of X. Because of this we define a function $f_X : T \to [0, 1]$ given by

$$f_X(t) = P(X = t)$$

referred to as a "probability mass function". Then for any event $A \subset T$ the quantity $P(X \in A)$ may be computed via

$$P(X \in A) = \sum_{t \in A} f_X(t) = \sum_{t \in A} P(X = t).$$

The function from Example 3.1.4 is a discrete random variable because it takes on one of the real values 0, 20, or 200. We calculated its probability mass function when describing its distribution and it is given by

$$f_X(0) = \frac{972}{1000}, f_X(20) = \frac{27}{1000}, f_X(200) = \frac{1}{1000}.$$

The function from Example 3.1.3 is not a discrete random variable by the above definition because its range is a collection of colors, not real numbers.

3.1.1 *Common Distributions*

When studying random variables it is often more important to know how they distribute probability onto their range than how they actually act as functions on their domains. As such it is useful to have a notation that recognizes the fact that two functions may be very different in terms of where they map domain elements, but nevertheless have the same range and produce the same distribution on this range.

DEFINITION 3.1.6. *Let $X : S \to T$ and $Y : S \to T$ be discrete random variables. We say X and Y have equal distriubtion provided $P(X = t) = P(Y = t)$ for all $t \in T$.*

There are many distributions which appear frequently enough they deserve their own special names for easy identification. We shall use the symbol $\sim$ to mean "is distributed as" or "is equal in distribution to". For example, in the definition below $X \sim Bernoulli(p)$ should be read as "X has a Bernoulli(p) distribution". This says nothing explicit about how X behaves as a function on its domain, but completely describes how X distributes probability onto its range.

DEFINITION 3.1.7. *The following are common discrete distributions which we have seen arise previously in the text.*

(a) $X \sim$ **Uniform**$(\{1, 2, \ldots, n\})$*: Let $n \geq 1$ be an integer. If X is a random variable such that $P(X = k) = \frac{1}{n}$ for all $1 \leq k \leq n$ then we say that X is a uniform random variable on the set $\{1, 2, \ldots, n\}$.*

(b) $X \sim$ **Bernoulli**(p)*: Let $0 \leq p \leq 1$. When X is a random variable such that $P(X = 1) = p$ and $P(X = 0) = 1 - p$ we say that X is a Bernoulli random variable with parameter p. This takes the concept of a "Bernoulli trial" which we have previously discussed and puts it in the context of a random variable where 1 corresponds to success and 0 corresponds to failure.*

(c) $X \sim$ **Binomial**(n,p): *Let $0 \leq p \leq 1$ and let $n \geq 1$ be an integer. If X is a random variable taking values in $\{0,1,\ldots,n\}$ having a probability mass function*

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for all $0 \leq k \leq n$, then X is a binomial random variable with parameters n and p. We have seen that such a quantity describes the number of successes in n Bernoulli trials.

(d) $X \sim$ **Geometric**(p): *Let $0 < p < 1$. If X is a random variable with values in $\{1,2,3,\ldots\}$ and a probability mass function*

$$P(X=k) = p \cdot (1-p)^{k-1}$$

for all $k \geq 1$, then X is a geometric random variable with parameter p. Such a random variable arises when determining how many Bernoulli trials must be attempted before seeing the first success.

(e) $X \sim$ **Negative Binomial** (r,p): *Let $0 < p < 1$. If X is a random variable with values in $\{r, r+1, r+2, \ldots\}$ and a probability mass function*

$$P(X=k) = \binom{k-1}{r-1} p^r \cdot (1-p)^{k-r}$$

for all $k \geq r$, then X is a negative binomial random variable with parameters (r,p). Such a random variable arises when determining how many Bernoulli trials must be attempted before seeing r successes.

(f) $X \sim$ **Poisson**(λ): *Let $\lambda > 0$. When X is a random variable with values in $\{0,1,2,\ldots\}$ such that its probability mass function is*

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for all $k \geq 0$, then X is called a Poisson random variable with parameter λ. We first used these distributions as approximations to a Binomial(n,p) when n was large and p was small.

(g) $X \sim$ **HyperGeo**(N,r,m): *Let N, r, and m be positive integers for which $r < N$ and $m < N$. Let X be a random variable taking values in the integers between $\min\{m,r\}$ and $\max\{0, m-(N-r)\}$ inclusive with probability mass function*

$$P(X=k) = \frac{\binom{r}{k}\binom{N-r}{m-k}}{\binom{N}{m}}$$

The random variable X is called hypergeoemtric with parameters N, r, and m. Such quantities occur when sampling without replacement.

EXERCISES

Ex. 3.1.1. Consider the experiment of flipping a coin four times and recording the sequence of heads and tails. Let S be the sample space of all sixteen possible orderings of the results. Let X be the function on S describing the number of tails among the flips. Let Y be the function on S describing the first flip (if any) to come up tails.

(a) Create a table as in Example 3.1.1 describing functions X and Y.

(b) Use the table to calculate $P(X = 2)$.

(c) Use the table to calculate $P(Y = 3)$.

Ex. 3.1.2. A pair of fair dice are thrown. Let X represent the larger of the two values on the dice and let Y represent the smaller of the two values.

(a) Describe S, the domain of functions X and Y. How many elements are in S?

(b) What are the ranges of X and Y. Do X and Y have the same range? Why or why not?

(c) Describe the distribution of X and describe the distribution of Y by finding the probability mass function of each. Is it true that X and Y have the same distribution ?

Ex. 3.1.3. A pair of fair dice are thrown. Let X represent the number of the first die and let Y represent the number of the second die.

(a) Describe S, the domain of functions X and Y. How many elements are in S?

(b) Describe T, the range of functions X and Y. How many elements are in T?

(c) Describe the distribution of X and describe the distribution of Y by finding the probability mass function of each. Is it true that X and Y have the same distribution ?

(d) Are X and Y the same function? Why or why not?

Ex. 3.1.4. Use the $\sim$ notation to classify the distributions of the random variables described by the scenarios below. For instance, if a scenario said, "let X be the number of heads in three flips of a coin" the approrpriate answer would be $X \sim \text{Binomial}(3, \frac{1}{2})$ since that describes the number of successes in three Bernoulli trials.

(a) Let X be the number of 5's seen in four die rolls. What is the distribution of X?

(b) Each ticket in a certain lottery has a 20% chance to be a prize-winning ticket. Let Y be the number of tickets that need to be purchased before seeing the first prize-winner. What is the distribution of Y?

(c) A class of ten students is comprised of seven women and three men. Four students are randomly selected from the class. Let Z denote the number of men among the four randomly selected students. What is the distribution of Z?

Ex. 3.1.5. Suppose X and Y are random variables.

(a) Explain why $X + Y$ is a random variable.

(b) Theorem 3.1.2 does not require that X be real-valued. Why do you suppose that our definition of "random variable" insisted that such functions should be real-valued?

Ex. 3.1.6. Let $X : S \to T$ be a discrete random variable. Suppose $\{B_i\}_{i\geq 1}$ are sequence of events in T then show that $X^{-1}(\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} X^{-1}(B_i)$ and that if B_i and B_j are disjoint, then so are $X^{-1}(B_i)$ and $X^{-1}(B_j)$.

3.2 INDEPENDENT AND DEPENDENT VARIABLES

Most interesting problems require the consideration of several different random variables and an analysis of the relationships among them. We have already discussed what it means for a collection of events to be independent and it is useful to extend this notion to random variables as well. As with events we will first describe the notion of pairwise independence of two objects before defining mutual independence of an arbitrary connection of objects.

3.2.1 *Independent Variables*

DEFINITION 3.2.1. **(Independence of a Pair of Random Variables)** *Two random variables X and Y are independent if $(X \in A)$ and $(Y \in B)$ are independent for every event A in the range of X and every event B in the range of Y.*

As events become more complicated and involve multiple random variables, a notational shorthand will become useful. It is common in probability to write $(X \in A, Y \in B)$ for the event $(X \in A) \cap (Y \in B)$ and we will begin using this convention at this point.

Further, even though the definition of $X : S \to T$ and $Y : S \to U$ being independent random variables requires that $(X \in A)$ and $(Y \in B)$ be independent for all events $A \subset T$ and $B \subset U$, for discrete random variables it is enough to verify the events $(X = t)$ and $(Y = u)$ are independent events for all $t \in T$ and $u \in U$ to conclude they are independent (See Exercise 3.2.12).

EXAMPLE 3.2.2. When we originally considered the example of rolling a pair of dice, we viewed the results as thirty-six equally likely outcomes. However, it is also possible to view the result of each die as a random variable in its own right, and then consider the possible results of the pair of random variables. Let $X, Y \sim \text{Uniform}(\{1, 2, 3, 4, 5, 6\})$ and suppose X and Y are independent. If $x, y \in \{1, 2, 3, 4, 5, 6\}$ what is $P(X = x, Y = y)$?

By indpendence $P(X = x, Y = y) = P(X = x)P(Y = y) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Therefore, the result is identical to the original perspective – each of the thirty-six outcomes of the pair of dice is equally likely. ■

DEFINITION 3.2.3. **(Mutual Indpendence of Random Variables)** *A finite collection of random variables $X_1, X_2, \ldots, X_n$ is mutually independent if the sets $(X_j \in A_j)$ are mutually independent for all events A_j in the ranges of the corresponding X_j.*

An arbitrary collection of random variables X_t where $t \in I$ for some index set I is mutually independent if every finite sub-collection is mutually independent.

For many problems it is useful to think about repeating a single experiment many times with the results of each repetition being independent from every other. Though the results are assumed

to be independent, the experiment itself remains the same, so the random variables produced all have the same distribution. The resulting sequence of random variables $X_1, X_2, X_3, \ldots$ is referred to as " i.i.d." (standing for "independent and identically distributed"). When considering such sequences we will sometimes write $X_1, X_2, X_3, \ldots$ are i.i.d. with distribution X, where X is a random variable that shares their common distribution.

EXAMPLE 3.2.4. Let $X_1, \ldots, X_n$ be i.i.d. with a Geometric(p) distribution. What is the probabilty that all of these random variables are larger than some positive integer j?

As a preliminary calculation, if $X \sim$ Geometric(p) and if $j \geq 1$ is an integer we may determine $P(X > j)$.

$$\begin{aligned} P(X > j) &= \sum_{i=j+1}^{\infty} P(X = i) \\ &= \sum_{i=j+1}^{\infty} p(1-p)^{i-1} \\ &= \frac{p \cdot (1-p)^j}{1-(1-p)} \\ &= (1-p)^j \end{aligned}$$

But each of X_1, X_2, $\ldots$, X_n have this distribution, so using the computatation above, together with independence,

$$\begin{aligned} P(X_1 > j, X_2 > j, \ldots, X_n > j) &= P(X_1 > j)P(X_2 > j)\ldots P(X_n > j) \\ &= (1-p)^j \cdot (1-p)^j \cdot \cdots \cdot (1-p)^j \\ &= (1-p)^{nj} \end{aligned}$$

■

3.2.2 *Conditional, Joint, and Marginal Distributions*

Consider a problem involving two random variables. Let X be the number of centimeters of rainfall in a certain forest in a given year, and let Y be the number of square meters of the forest burned by fires that same year. It seems these variables should be related; knowing one should affect the probabilities associated with the values of the other. Such random variables are not independent of each other and we now introduce several ways to compute probabilities under such circumstances. An important concept toward this end is the notion of a "conditional distribution" which reflects the fact that the occurrence of an event may affect the likely values of a random variable.

DEFINITION 3.2.5. *Let X be a random variable on a sample space S and let $A \subset S$ be an event such that $P(A) > 0$. Then the probability Q described by*

$$Q(B) = P(X \in B|A) \tag{3.2.1}$$

is called the "conditional distribution" of X given the event A.

As with any discrete random variable, the distribution is completely determined by the probabilities associated with each possible value the random variable may assume. This means the conditional distribution may be considered known provided the values of $P(X = a|A)$ are known for every $a \in \text{Range}(X)$. Though this definition allows for A to be any sort of event, in this section we will mainly consider examples where A describes the outcome of some random variable. So a notation like $P(X|Y = b)$ will be the conditional distribution of the random variable X given that the variable Y is known to have the value b.

In many cases random variables are dependent in such a way that the distribution of one variable is known in terms of the values taken on by another.

EXAMPLE 3.2.6. Let $X \sim \text{Uniform}(\{1, 2\})$ and let Y be the number of heads in X tosses of a fair coin. Clearly X and Y should not be independent. In particular, a result of $Y = 0$ could occur regardless of the value of X, but a result of $Y = 2$ guarantees that $X = 2$ since two heads could not be observed with just one flip on the coin. Any information regarding X or Y may influence the distribution of the other, but the description of the variables makes it clearest how Y depends on X. If $X = 1$ then Y is the number of heads in one flip of a fair coin. Letting A be the event $(X = 1)$ and using the terminology of (3.2.1) from Definition 3.2.5, we can say the conditional distribution of Y given that $X = 1$ is a Bernoulli($\frac{1}{2}$). We will use the notation

$$(Y|X = 1) \sim Bernoulli(\frac{1}{2})$$

to indicate this fact. In other words, this notation means the same thing as the pair of equations

$$\begin{aligned} P(Y = 0|X = 1) &= \frac{1}{2} \\ P(Y = 1|X = 1) &= \frac{1}{2} \end{aligned}$$

If $X = 2$ then Y is the number of heads in two flips of a fair coin and therefore $(Y|X = 2) \sim \text{Binomial}(2, \frac{1}{2})$ which means the following three equations hold:

$$\begin{aligned} P(Y = 0|X = 2) &= \frac{1}{4} \\ P(Y = 1|X = 2) &= \frac{1}{2} \\ P(Y = 2|X = 2) &= \frac{1}{4} \end{aligned}$$

■

The conditional probabilities of the previous example were easily determined in part because the description of Y was already given in terms of X, but frequently random variables may be dependent in some way that is not so explicitly described. A more general method of expressing the dependence of two (or more) variables is to present the probabilities associated with all combinations of possible values for every variable. This is known as their joint distribution.

DEFINITION 3.2.7. *If X and Y are discrete random variables, the "joint distribution" of X and Y is the probability Q on pairs of values in the ranges of X and Y defined by*

$$Q((a, b)) = P(X = a, Y = b).$$

The definition may be expanded to a finite collection of discrete random variables X_1, X_2, ..., X_n for which the joint distribution of all n variables is the probability defined by

$$Q((a_1, a_2, \ldots, a_n)) = P(X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n).$$

In the above definition as discussed before for any event D,

$$Q(D) = \sum_{(a_1,a_2,\ldots,a_n)\in D} Q((a_1, a_2, \ldots, a_n)).$$

For a pair of random variables with few possible outcomes, it is common to describe the joint distribution using a chart for which the columns correspond to possible X values, the rows to possible Y values, and for which the entries of the chart are probabilities.

EXAMPLE 3.2.8. Let X and Y be the dependent variables described in Example 3.2.6. The X variable will be either 1 or 2. The Y variable could be as low as 0 (if no heads are flipped) or as high as 2 (if two coins are flipped and both show heads). Since $\text{Range}(X) = \{1, 2\}$ and since $\text{Range}(Y) = \{0, 1, 2\}$, the pair (X, Y) could potentially be any of the six possible pairings (though, in fact, one of the pairings has probability zero). To find the joint distribution of X and Y we must calculate the probabilities of each possibility. In this case the values may be obtained using the definition of conditional probability. For instance,

$$P(X = 1, Y = 0) = P(Y = 0|X = 1) \cdot P(X = 1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

and

$$P(X = 1, Y = 2) = P(Y = 2|X = 1) \cdot P(X = 1) = 0 \cdot \frac{1}{2} = 0$$

The entire joint distribution $P(X = a, Y = b)$ is described by the following chart.

	$X = 1$	$X = 2$
$Y = 0$	1/4	1/8
$Y = 1$	1/4	1/4
$Y = 2$	0	1/8

■

Knowing the joint distribution of random variables gives a complete picture of the probabilities associated with those variables. From that information it is possible to compute all conditional probabilities of one variable from another. For instance, in the example analyzed above, the variable Y was originally described in terms of how it depended on X. However, this also means that X should be dependent on Y. The joint distribution may be used to determine how.

EXAMPLE 3.2.9. Let X and Y be the variables of Example 3.2.6. How may the conditional distributions of X given values of Y be determined?

There will be three different conditional distributions depending on whether $Y = 0$, $Y = 1$, or $Y = 2$. Below we will solve the $Y = 0$ case. The other two cases will be left as exercises. The

conditional distribution of X given that $Y = 0$ is determined by the values of $P(X = 1|Y = 0)$ and $P(X = 2|Y = 0)$ both of which may be computed using Bayes' rule.

$$\begin{aligned} P(X = 1|Y = 0) &= \frac{P(Y = 0|X = 1) \cdot P(X = 1)}{P(Y = 0)} \\ &= \frac{P(Y = 0|X = 1) \cdot P(X = 1)}{P(Y = 0|X = 1) \cdot P(X = 1) + P(Y = 0|X = 2) \cdot P(X = 2)} \\ &= \frac{(1/2)(1/2)}{(1/2)(1/2) + (1/4)(1/2)} = \frac{2}{3} \end{aligned}$$

Since the only values for X are 1 and 2 it must be that $P(X = 2|Y = 0) = \frac{1}{3}$. ■

Just because X and Y are dependent on each other doesn't mean they need to be thought of as a pair. It still makes sense to talk about the distribution of X as a random variable in its own right while ignoring its dependence on the variable Y. When there are two or more variables under discussion, the distribution of X alone is sometimes called the "marginal" distribution of X because it can be computed using the margins of the chart describing the joint distribution of X and Y.

EXAMPLE 3.2.10. Continue with X and Y as described in Example 3.2.6. Below is the chart describing the joint distribution of X and Y that was created in Example 3.2.8, but with the addition of one column on the right and one row at the bottom. The entries in the extra column are the sums of the values in the corresponding row; likewise the entries in the extra row are the sums of the values in the corresponding column.

	$X = 1$	$X = 2$	Sum
$Y = 0$	1/4	1/8	3/8
$Y = 1$	1/4	1/4	4/8
$Y = 2$	0	1/8	1/8
Sum	1/2	1/2	

The values in the right hand margin (column) exactly describe the distribution of Y. For instance the event $(Y = 0)$ can be partitioned into two disjoint events $(X = 1, Y = 0) \cup (X = 2, Y = 0)$ each of which is already described in the joint distribution chart. Adding them together gives the result that $P(Y = 0) = \frac{3}{8}$. In a similar fashion, the bottom margin (row) describes the distribution of X. This extended chart also makes it numerically clearer why these two random variables are dependent. For instance,

$$P(X = 1, Y = 0) = \frac{1}{4} \qquad \text{while} \qquad P(X = 1) \cdot P(Y = 0) = \frac{3}{16}$$

Since these quantities are unequal, the random variables cannot be independent. ■

In general, knowing the marginal distributions of X and Y is not sufficient information to reconstruct their joint distribution. This is because the marginal distributions do not provide any information about how the random variables relate to each other. However, if X and Y happen to be independent, then their joint distribution may easily be computed from the marginals since

$$P(X = x, Y = y) = P(X = x)P(Y = x)$$

3.2.3 *Memoryless Property of the Geometric Random Variable*

It is also possible to calculate conditional probabilities of a random variable based on subsets of its own values. A particularly important example of this is the "memoryless property" of geometric random variables.

EXAMPLE 3.2.11. Suppose we toss a fair coin until the first head appears. Let X be the number of tosses performed. We have seen in Example 2.1.2 that $X \sim \text{Geometric}(\frac{1}{2})$. Note that if m is a positive integer,

$$P(X > m) = \sum_{k=m+1}^{\infty} P(X = k) = \sum_{k=m+1}^{\infty} \frac{1}{2^k} = \frac{1}{2^m}$$

Now let n be a positive integer and suppose we take the event $(X > n)$ as given. In other words, we assume we know that none of the first n flips resulted in heads. What is the conditional distribution of X given this new information? A routine calculation shows

$$P(X > n + m | X > n) = \frac{P(X > n + m)}{P(X > n)} = \frac{\frac{1}{2^{m+n}}}{\frac{1}{2^n}} = \frac{1}{2^m}$$

As a consequence,

$$P(X > n + m | X > n) = P(X > m). \tag{3.2.2}$$

Given that a result of heads has not occurred by the n^{th} flip, the probability that such a result will require at least m more flips is identical to the (non-conditional) probability the result would have required more than m flips from the start. In other words, if we know that the first n flips have not yet produced a head, the number of additional flips required to observe the first head still is a $\text{Geometric}(\frac{1}{2})$ random variable. This is called the "memoryless property" of the geometric distribution since it can be interpreted to mean that when waiting times are geometrically distributed, no matter how long we wait for an event to occur, the future waiting time always looks the same given that the event has not occurred yet. The result remains true of geometric variables of any parameter p, a fact which we leave as an exercise. ■

3.2.4 *Multinomial Distributions*

Consider a situation similar to that of Bernoulli trials, but instead of results of each attempt limited to success or failure, suppose there are many different possible results for each trial. As with the Bernoulli trial cases we assume that the trials are mutually independent, but identically distributied. In the next example we will show how to calculate the joint distribution for the random variables representing the number of times each outcome occurs.

EXAMPLE 3.2.12. Suppose we perform n i.i.d. trials each of which has k different possible outcomes. For $j = 1, 2, \ldots, k$, let p_j represent the probability any given trial results in the j^{th} outcome and let X_j represent the number of the n trials that result in the j^{th} outcome. The joint distribution of all of the random variables X_1, X_2, …, X_k is called a "multinomial distribution".

Let $B(x_1, x_2, \ldots, x_k) = \{X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\}$. Then,

$$P(B(x_1, x_2, \ldots, x_k)) = \sum_{\omega \in B(x_1, x_2, \ldots, x_k)} P(\{\omega\})$$

Each $\omega \in B(x_1, x_2, \ldots, x_k)$ is an element in the sample space corresponding to the j^{th} outcome occuring exactly x_j times. Since the trials are independent, and since an outcome j occurs in x_j trials, each of which had probability p_j, this means

$$P(\{\omega\}) = \prod_{j=1}^{k} p_j^{x_j}$$

Consequently, each outcome in $B(x_1, x_2, \ldots, x_k)$ has the same probability. So to determine the likelihood of the event, we need only determine $|B(x_1, x_2, \ldots, x_k)|$, the number of outcomes the event contains. The calculation of this quantity is a combinatorial problem; it is the number of ways of allocating n balls in k boxes, such that x_j of them fall into box j. We leave it as an exercise to prove that

$$|B(x_1, x_2, \ldots, x_k)| = \frac{n!}{x_1! \, x_2! \, \ldots \, x_k!}$$

With that computation complete, the joint distribution of $X_1, X_2, \ldots X_k$ is given by

$$P(X_1 = x_1, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1! \, x_2! \ldots x_k!} \prod_{j=1}^{k} p_j^{x_j} & \text{if } x_j \in \{0, 1, \ldots, n\} \\ & \text{and } \sum_{j=1}^{k} x_j = n \\ 0 & \text{otherwise} \end{cases}$$

EXERCISES

Ex. 3.2.1. An urn has four balls labeled $1, 2, 3$, and 4. A first ball is drawn and its number is denoted by X. A second ball is then drawn from the three remaining balls in the urn and its number is denoted by Y.

(a) Calculate $P(X = 1)$.

(b) Calculate $P(Y = 2|X = 1)$.

(c) Calculate $P(Y = 2)$.

(d) Calculate $P(X = 1, Y = 2)$.

(e) Are X and Y independent? Why or why not?

Ex. 3.2.2. Two dice are rolled. Let X denote the sum of the dice and let Y denote the value of the first die.

(a) Calculate $P(X = 7)$ and $P(Y = 4)$.

(b) Calculate $P(X = 7, Y = 4)$.

(c) Calculate $P(X = 5)$ and $P(Y = 4)$.

(d) Calculate $P(X = 5, Y = 4)$.

(e) Are X and Y independent? Why or why not?

Ex. 3.2.3. Let X and Y be the variables described in Example 3.2.6.

(a) Determine the conditional distribution of X given that $Y = 1$.

(b) Determine the conditional distribution of X given that $Y = 2$.

Ex. 3.2.4. Let X and Y be random variables with joint distribution given by the chart below.

	$X=0$	$X=1$	$X=2$
$Y=0$	1/12	0	3/12
$Y=1$	2/12	1/12	0
$Y=2$	3/12	1/12	1/12

(a) Compute the marginal distributions of X and Y.

(b) Compute the conditional distribution of X given that $Y=2$.

(c) Compute the conditional distribution of Y given that $X=2$.

(d) Carry out a computation to show that X and Y are not independent.

Ex. 3.2.5. Let X be a random variable with range $\{0,1\}$ and distribution

$$P(X=0)=\frac{1}{3} \qquad \text{and} \qquad P(X=1)=\frac{2}{3}$$

and let Y be a random variable with range $\{0,1,2\}$ and distribution

$$P(Y=0)=\frac{1}{5}, \qquad P(Y=1)=\frac{1}{5}, \qquad \text{and} \qquad P(Y=2)=\frac{3}{5}$$

Suppose that X and Y are independent. Create a chart describing the joint distribution of X and Y.

Ex. 3.2.6. Consider six independent trials each of which are equally likely to produce a result of 1, 2, or 3. Let X_j denote the number of trials that result in j. Calculate $P(X_1=1, X_2=2, X_3=3)$.

Ex. 3.2.7. Prove the combinatorial fact from Example 3.2.12 in the following way. Let $A_n(x_1,x_2,\ldots,x_k)$ denote the number of ways of putting n balls into k boxes in such a way that exactly x_j balls wind up in box j for $j=1,2,\ldots,k$.

(a) Prove that $A_n(x_1,x_2,\ldots,x_k)=\binom{n}{x_1}A_{n-x_1}(x_2,x_3,\ldots,x_k)$.

(b) Use part (a) and induction to prove that $A_n(x_1,x_2,\ldots,x_k)=\frac{n!}{x_1!\,x_2!\ldots x_k!}$.

Ex. 3.2.8. Let X be the result of a fair die roll and let Y be the number of heads in X coin flips.

(a) Both X and $(Y|X=n)$ can be written in terms of common distributions using the $\sim$ notation. What is the distribution of X? What is the distribution of $(Y|X=n)$ for $n=1,\ldots 6$?

(b) Determine the joint distribution for X and Y.

(c) Calculate the marginal distribution of Y.

(d) Compute the conditional distribution of X given that $Y=6$.

(e) Compute the conditional distribution of X given that $Y=0$.

(f) Perform a computation to prove that X and Y are not independent.

Ex. 3.2.9. Suppose the number of earthquakes that occur in a year, anywhere in the world, is a Poisson random variable with mean λ. Suppose the probability that any given earthquake has magnitude at least 5 on the Richter scale is p independent of all other quakes. Let $N\sim\text{Poisson}(\lambda)$ be the number of earthquakes in a year and let M be the number of earthquakes in a year with magnitude at least 5, so that $(M|N=n)\sim\text{Binomial}(n,p)$.

(a) Calculate the joint distribution of M and N.

(b) Show that the marginal distribution of M is determined by

$$P(M = m) = \frac{1}{m!}e^{-\lambda}(\lambda p)^m \sum_{n=m}^{\infty} \frac{\lambda^{n-m}}{(n-m)!}(1-p)^{n-m}$$

for $m > 0$.

(c) Perform a change of variables (where $k = n - m$) in the infinite series from part (b) to prove

$$P(M = m) = \frac{1}{m!}e^{-\lambda}(\lambda p)^m \sum_{k=0}^{\infty} \frac{(\lambda(1-p))^k}{k!}$$

(d) Use part (c) together with the infinite series equality $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ to conclude that $M \sim$ Poisson(λp).

Ex. 3.2.10. Let X be a discrete random variable which has $\mathbb{N} = \{1, 2, 3, \dots\}$ as its range. Suppose that for all positive integers m and n, X has the memoryless property – $P(X > n + m | X > n) = P(X > m)$. Prove that X must be a geometric random variable. [Hint: Define $p = P(X = 1)$ and use the memoryless property to calculate $P(X = n)$ inductively].

Ex. 3.2.11. A discrete random variable X is called "constant" if there is a single value c for which $P(X = c) = 1$.

(a) Prove that if X is a constant discrete random variable then X is independent of itself.

(b) Prove that if X is a discrete random variable which is independent of itself, then X must be constant. [Hint: It may help to look at Exercise 1.4.8].

Ex. 3.2.12. Let $X : S \to T$ and $Y : S \to U$ be discrete random variables. Show that if

$$P(X = t, Y = u) = P(X = t)P(Y = u)$$

for all $t \in T$ and $u \in U$ then X and Y are independent random variables.

3.3 FUNCTIONS OF RANDOM VARIABLES

There are many circumstances where we want to consider functions applied to random variables as inputs of functions. For a simple geometric example, suppose a rectangle is selected in such a way that its width X and its length Y are both random variables with known joint distribution. The area of the rectangle is $A = XY$ and since X and Y are random, it should be that A is random as well. How may the distribution of A be calculated from the joint distribution of X and Y? In general, if a new random variable Z depends on random variables X_1, X_2, ..., X_n which have a given joint distribution, how may the distribution of Z be calculated from what is already known? In this section we discuss the answers to such questions and also address related issues surrounding independence.

If $X : S \to T$ is a random variable and if $f : T \to R$ is a function, then the quantity $f(X)$ makes sense as a composition of functions $f \circ X : S \to R$. In fact, since $f(X)$ is defined on the sample space S, this new composition is itself a random variable.

The same reasoning holds for functions of more than one variable. If $X_1, X_2, \ldots, X_n$ are random variables then $f(X_1, X_2, \ldots, X_n)$ is a random variable provided f is defined for the values the X_j variables produce. Below we illustrate how to calculate the distribution of $f(X_1, X_2, \ldots, X_n)$ in terms of the joint distribution of the X_j input variables. We demonstrate the method with several examples followed by a general theorem.

3.3.1 *Distribution of $f(X)$ and $f(X_1, X_2, \ldots, X_n)$*

The distribution of $f(X)$ involves the probability of events such as $(f(X) = a)$ for values of a that the function may produce. The key to calculating this probability is that these events may be rewritten in terms of the input values of X instead of the output values of $f(X)$.

EXAMPLE 3.3.1. Let $X \sim \text{Uniform}(\{-2, -1, 0, 1, 2\})$ and let $f(x) = x^2$. Determine the range and distribution of $f(X)$.

Since $f(X) = X^2$, the values that $f(X)$ produces are the squares of the values that X produces. Squaring the values in $\{-2, -1, 0, 1, 2\}$ shows the range of $f(X)$ is $\{0, 1, 4\}$. The probabilities that $f(X)$ takes on each of these three values determine the distribution of $f(X)$ and these probabilities can be easily calculated from the known probabilities associated with X.

$$\begin{aligned} P(f(X) = 0) &= P(X = 0) = \frac{1}{5} \\ P(f(X) = 1) &= P((X = 1) \cup (X = -1)) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5} \\ P(f(X) = 4) &= P((X = 2) \cup (X = -2)) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5} \end{aligned}$$

■

A complication with this method is that there may be many different inputs that produce the same output. Sometimes a problem requires careful consideration of all ways that a given output may be produced. For instance,

EXAMPLE 3.3.2. What is the probability the sum of three dice will equal six? Let X, Y, and Z be the results of the first, second, and third die respectively. These are i.i.d. random variables each distributed as $\text{Uniform}(\{1, 2, 3, 4, 5, 6\})$. A sum of six can be arrived at in three distinct ways:

Case I: through three rolls of 2;
Case II: through one roll of 3, one roll of 2, and one roll of 1; or
Case III: through one roll of 4 and two rolls of 1

The first of these is the simplest to deal with since independence gives

$$P(X = 2, Y = 2, Z = 2) = P(X = 2) \cdot P(Y = 2) \cdot P(Z = 2) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{216}$$

The other cases involve a similar computation, but are complicated by the consideration of which number shows up on which die. For instance, both events $(X = 1, Y = 2, Z = 3)$ and $(X = 3, Y = 2, Z = 1)$ are included as part of Case II as are four other permutations of the numbers. Likewise Case III includes three permutations, one of which is $(X = 4, Y = 1, Z = 1)$. Putting all three cases together,

$$\begin{aligned} P(\text{sum of 6}) &= P(\text{Case I}) + P(\text{Case II}) + P(\text{Case III}) \\ &= \frac{1}{216} + 6(\frac{1}{216}) + 3(\frac{1}{216}) \\ &= \frac{5}{108} \end{aligned}$$

So there is slightly less than a 5% chance three rolled dice will produce a sum of six. ■

This method may also be used to show relationships among the common (named) distributions that have been previously described, as in the next two examples.

EXAMPLE 3.3.3. Let $X, Y \sim \text{Bernoulli}(p)$ be two independent random variables. If $Z = X + Y$, show that $Z \sim \text{Binomial}(2, p)$.

This result should not be surpirsing given how Bernoulli and binomial distributions arose in the first place. Each of X and Y produces a value of 0 if the corresponding Bernoulli trial was a failure and 1 if the trial was a success. Therefore $Z = X + Y$ equals the total number of successes in two independent Bernoulli trials, which is exactly what led us to the binomial distribution in the first place. However, it is instructive to consider how this problem relates to the current topic of discussion.

Since each of X and Y is either 0 or 1 the possible values of Z are in the set $\{0,1,2\}$. A result of $Z = 0$ can only occur if both X and Y are zero. So,

$$\begin{aligned} P(Z = 0) &= P(X = 0, Y = 0) \\ &= P(X = 0) \cdot P(Y = 0) \\ &= (1-p)(1-p) = (1-p)^2 \end{aligned}$$

Similarly, $P(Z = 2) = P(X = 1) \cdot P(Y = 1) = p^2$.

There are two different ways that Z could equal 1, either $X = 1$ and $Y = 0$, or $X = 0$ and $Y = 1$. So,

$$\begin{aligned} P(Z = 1) &= P((X = 1, Y = 0) \cup (X = 0, Y = 1)) \\ &= P(X = 1, Y = 0) + P(X = 0, Y = 1) \\ &= p(1-p) + (1-p)p = 2p(1-p) \end{aligned}$$

These values of $P(Z = 0), P(Z = 1)$, and $P(Z = 2)$ are exactly what define $Z \sim \text{Binomial}(2, p)$. ■

Two of the previous three examples involving adding random variables together. In fact, addition is one of the most common examples of applying functions to random quantities. In the previous situations, calculating the distribution of the sum was relatively simple because the component variables only had finitely many outcomes. But now suppose X and Y are random varialbes taking values in $\{0, 1, 2, \dots\}$ and suppose $Z = X + Y$. How could $P(Z = n)$ be calculated?

Since both X and Y are non-negative and since $Z = X + Y$, the value of Z must be at least as large as either X or Y individually. If $Z = n$, then X could take on any value $j \in \{0, 1, \dots, n\}$, but once that value is determed, the value of Y is compelled to be $n - j$ to give the appropriate sum. In other words, the event $(Z = n)$ partitions into the following union.

$$(Z = n) = \bigcup_{j=0}^{n} (X = j, Y = n - j)$$

When X and Y are independent, this means

$$\begin{aligned} P(Z=n) &= P(\bigcup_{j=0}^{n}(X=j, Y=n-j)) \\ &= \sum_{j=0}^{n} P(X=j, Y=n-j) \\ &= \sum_{j=0}^{n} P(X=j)\cdot P(Y=n-j) \end{aligned}$$

Such a computation is usually referred to as a "convolution" which will be addressed more generally later in the text. It occurs regularly when determining the distribution of sums of independent random variables.

EXAMPLE 3.3.4. Let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ be independent random variables.

(a) Let $Z = X + Y$. Find the distribution of Z.

(b) Find the conditional distribution of $X \mid Z$.

Now

$$P(X=x, Y=y) = P(X=x)\cdot P(Y=y) = e^{-\lambda_1}\frac{\lambda_1^x}{x!}\cdot e^{-\lambda_2}\frac{\lambda_2^y}{y!} \qquad \text{for} \quad x, y \in \{0,1,2,\dots\}$$

(a) As computed above the distribution of Z is given by the convolution. For any $n = 0, 1, 2, \dots$ we have

$$\begin{aligned} P(Z=n) &= P(X+Y=n) \\ &= \sum_{j=0}^{n} P(X=j)\cdot P(Y=n-j) \\ &= \sum_{j=0}^{n} e^{-\lambda_1}\frac{\lambda_1^j}{j!}\cdot e^{-\lambda_2}\frac{\lambda_2^{n-j}}{(n-j)!} \\ &= e^{-(\lambda_1+\lambda_2)}\sum_{j=0}^{n}\frac{\lambda_1^j\lambda_2^{n-j}}{j!(n-j)!} \\ &= e^{-(\lambda_1+\lambda_2)}\frac{1}{n!}\sum_{j=0}^{n}\frac{n!}{j!(n-j)!}\lambda_1^j\lambda_2^{n-j} \\ &= e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!} \end{aligned}$$

where in the last line we have used the binomial expansion (2.1.1). Hence we can conclude that $Z \sim$ Poisson $(\lambda_1 + \lambda_2)$. The above calculation is easily extended by an induction argument to obtain the fact that if $\lambda_i > 0$, X_i, $1 \le i \le k$ are independent Poisson(λ_i) distributed random variables (respectively). Then $Z = \sum_{i=1}^{k} X_i$ has Poisson $(\sum_{i=1}^{k}\lambda_i)$ distribution. Thus if we have k independent Poisson (λ) random variables then $\sum_{i=1}^{k} X_i$ has Poisson$(k\lambda)$ distribution.

(b) We readily observe that X and Z are dependent. We shall now try to understand the conditional distribution of $(X|Z=n)$. Since the range of X and Y does not have any negative

numbers, given that $Z = X + Y = n$, X can only take values in $\{0, 1, 2, 3, \ldots, n\}$. For $k \in \{0, 1, 2, 3, \ldots, n\}$ we have

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

Hence $(X|Z = n) \sim \text{Binomial}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$.

■

The point of the examples above is that a probability associated with a functional value $f(X_1, X_2, \ldots, X_n)$ may be calculated directly from the probabilities associated with the input variables $X_1, X_2, \ldots, X_n$. The following theorem explains how this may be accomplished generally for any number of variables.

THEOREM 3.3.5. *Let X_1, X_2, ..., X_n be random variables defined on a single sample space S. Let f be a function of n variables for which $f(X_1, X_2, \ldots, X_n)$ is defined in the range of the X_j variables. Let B be a subset of the range of f. Then,*

$$P(f(X_1, X_2, \ldots, X_n) \in B) = P((X_1, X_2, \ldots, X_n) \in f^{-1}(B))$$

Proof - First note that both of the events $(f(X_1, X_2, \ldots, X_n) \in B)$ and $((X_1, X_2, \ldots, X_n) \in f^{-1}(B))$ are subsets of S since outcomes $s \in S$ determine the values of the X_j variables which in turn determine the output of f. The theorem follows immediately from the set theoretic fact that

$$f(X_1(s), X_2(s), \ldots, X_n(s)) \in B \iff (X_1(s), X_2(s), \ldots X_n(s)) \in f^{-1}(B)$$

This is because the expression $f(X_1(s), X_2(s), \ldots, X_n(s)) \in B$ is what defines s to be an outcome in the event $(f(X_1, X_2, \ldots, X_n) \in B)$. Likewise the expression $(X_1(s), X_2(s), \ldots X_n(s)) \in f^{-1}(B)$ defines s to in the event $((X_1, X_2, \ldots X_n) \in f^{-1}(B))$. Since these events are equal, they have the same probability. ■

3.3.2 *Functions and Independence*

If X and Y are independent random variables, does that guarantee that functions $f(X)$ and $g(Y)$ of these random variables are also indpendent? If we take the intuitive view of independence as saying "knowing information about X does not affect the probabilities associated with Y" then it seems the answer should be "yes". After all, X determines the value of $f(X)$ and Y determines the value of $g(Y)$. So information about $f(X)$ should translate to information about X and infromation about $f(Y)$ should translate to information about Y. Therefore if information about

$f(X)$ affected probabilities associated with $g(Y)$, then it seems there should be information about X that would affect the probability assoicated with Y. Below we generalize this argument and make it more rigorous.

THEOREM 3.3.6. *Let $n > 1$ be a positive integer. For each $j \in \{1, 2, \ldots, n\}$ define a positive integer m_j and suppose $X_{i,j}$ is an array of mutually independent random variables for $j \in \{1, 2, \ldots, n\}$ and $i \in \{1, 2, \ldots, m_j\}$. Let f_j be functions such that the quantity*

$$Y_j = f_j(X_{1,j}, X_{2,j}, \ldots X_{m_j,j})$$

is defined for the outputs of the $X_{i,j}$ variables. Then the resulting variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent.

Informally this theorem says that random quantities produced from independent inputs will, themselves, be independent.

Proof - Let $B_1, B_2, \ldots, B_n$ be sets in the ranges of $Y_1, Y_2, \ldots, Y_n$ respectively. Use of independence and set-theoretic identities then shows

$$\begin{aligned}
&P(Y_1 \in B_1, \ldots, Y_n \in B_n) \\
&= P(f_1(X_{1,1}, \ldots, X_{m_1,1}) \in B_1, \ldots, f_n(X_{1,n}, \ldots, X_{m_n,n}) \in B_n) \\
&= P((X_{1,1}, \ldots, X_{m_1,1}) \in f_1^{-1}(B_1), \ldots, (X_{1,n}, \ldots, X_{m_n,n}) \in f_n^{-1}(B_n)) \\
&= \prod_{i=1}^{n} P((X_{i,1}, \ldots, X_{m_i,i}) \in f_i^{-1}(B_i)) \\
&= \prod_{i=1}^{n} P(f_i(X_{i,1}, \ldots, X_{m_i,i}) \in B_i) \\
&= P(Y_1 \in B_1) \cdots P(Y_n \in B_n)
\end{aligned}$$

which proves that $Y_1, Y_2, \ldots, Y_n$ are mutually independent. ∎

EXERCISES

Ex. 3.3.1. Let $X \sim \text{Uniform}(\{1, 2, 3\})$ and $Y \sim \text{Uniform}(\{1, 2, 3\})$ be independent and let $Z = X + Y$.

(a) Determine the range of Z.

(b) Determine the distriubtion of Z.

(c) Is Z uniformly distributed over its range?

Ex. 3.3.2. Consider the experiment of rolling three dice and calculating the sum of the rolls. Answer the following questions.

(a) What is the range of possible results of this experiment?

(b) Calculate the probability the sum equals three.

(c) Calculate the probability the sum equals four.

(d) Calculate the probability the sum equals five.

(e) Calculate the probability the sum equals ten.

Ex. 3.3.3. Let $X \sim \text{Bernoulli}(p)$ and $Y \sim \text{Bernoulli}(q)$ be independent.

(a) Prove that XY is a Bernoulli random variable. What is its parameter?

(b) Prove that $(1 - X)$ is a Bernoulli random variable. What is its parameter?

(c) Prove that $X + Y - XY$ is a Bernoulli random variable. What is its parameter?

Ex. 3.3.4. Let $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$. Assume X and Y are independent and let $Z = X + Y$. Prove that $Z \sim \text{Binomial}(m + n, p)$.

Ex. 3.3.5. Let $X \sim \text{Negative Binomial}(r, p)$ and $Y \sim \text{Negative Binomial}(s, p)$. Assume X and Y are independent and let $Z = X + Y$. Prove that $Z \sim \text{Negative Binomial}(r + s, p)$.

Ex. 3.3.6. Consider one flip of a single fair coin. Let X denote the number of heads on the flip and let Y denote the number of tails on the flip.

(a) Show that $X, Y \sim \textit{Bernoulli}(\frac{1}{2})$.

(b) Let $Z = X + Y$ and explain why $P(Z = 1) = 1$.

(c) Since (b) clearly says that Z cannot be a $\textit{Binomial}(2, \frac{1}{2})$, explain why this result does not conflict with the conclusion of Example 3.3.3.

Ex. 3.3.7. Let $X \sim \text{Geometric}(p)$ and $Y \sim \text{Geometric}(p)$ be independent and let $Z = X + Y$.

(a) Determine the range of Z.

(b) Use a convolution to prove that $P(Z = n) = (n - 1)p^2(1 - p)^{n-2}$.

(c) Recall from the discussion of Geometric distributions that $(X = 1)$ is the most likely result for X and $(Y = 1)$ is the most likely result for Y. This does not imply that $(Z = 2)$ is the most likely outcome for Z. Determine the values of p for which $P(Z = 3)$ is larger than $P(Z = 2)$.

Ex. 3.3.8. Let X_1, X_2, X_3, X_4 be an i.i.d. sequence of Bernouli(p) random variables. Let $Y = X_1 + X_2 + X_3 + X_4$. Prove that $P(Y = 2) = 6p^2(1 - p)^2$.

Ex. 3.3.9. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of Bernoulli(p) random variables. Let $Y = X_1 + X_2 + \cdots + X_n$. Prove that $Y \sim \text{Binomial}(n, p)$.

Ex. 3.3.10. Let $X_1, X_2, \ldots, X_r$ be an i.i.d. sequence of Geometric (p) random variables. Let $Y = X_1 + X_2 + \cdots + X_r$. Prove that $Y \sim \text{Negative binomial}(r, p)$.

Ex. 3.3.11. Let X_1, X_2, X_3, X_4 be an i.i.d. sequence of Bernoulli(p) random variables. Let $Y = X_1 + X_2$ and let $Z = X_3 + X_4$. Note that Example 3.3.3 guarantees that $Y, Z \sim \text{Binomial}(2, p)$.

(a) Create a chart describing the joint distribution of Y and Z.

(b) Use the chart from (a) to explain why Y and Z are independent.

(c) Explain how you could use Theorem 3.3.6 to reach the conclusion that Y and Z are independent without calculating their joint distribution.

Ex. 3.3.12. Let X_1, X_2, X_3 be an i.i.d. sequence of Bernoulli(p) random variables. Let $Y = X_1 + X_2$ and let $Z = X_2 + X_3$. Note that Example 3.3.3 guarantees that $Y, Z \sim \text{Binomial}(2, p)$.

(a) Create a chart describing the joint distribution of Y and Z.

(b) Use the chart from (a) to explain why Y and Z are not independent.

(c) Explain why the conclusion from (b) is not inconsistant with Theorem 3.3.6.

Ex. 3.3.13. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of discrete random variables and let Z be the maximum of these n variables. Let r be a real number and let $R = P(X_1 \leq r)$. Prove that $P(Z \leq r) = R^n$.

Ex. 3.3.14. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of discrete random variables and let Z be the minimum of these n variables. Let r be a real number and let $R = P(X_1 \leq r)$. Prove that $P(Z \leq r) = 1 - (1 - R)^n$.

Ex. 3.3.15. Let $X \sim \text{Geometric}(p)$ and let $Y \sim \text{Geometric}(q)$ be independent random variables. Let Z be the smaller of X and Y. It is a fact that Z is also geometrically distributed. This problem asks you to prove this fact using two different methods.
METHOD I:

(a) Explain why the event $(Z = n)$ can be written as the disjoint union

$$(Z = n) = (X = n, Y = n) \cup (X = n, Y > n) \cup (X > n, Y = n)$$

(b) Recall from the proof of the memoryless property of geometric random variables that $P(X > m) = \frac{1}{2^m}$. Use this fact and part (a) to prove that

$$P(Z = n) = [(1-p)(1-q)]^{n-1}(pq + p(1-q) + (1-p)q)$$

(c) Use (b) to conclude that $Z \sim \text{Geometric}(r)$ for some quantity r and calculate the value of r in terms of the p and q.

METHOD II: Recall that geometric random variables first arose from noting the time it takes for a sequence of Bernoulli trials to first produce a success. With that in mind, let $A_1, A_2, \ldots$ be Bernoulli(p) random variables and let $B_1, B_2, \ldots$ be Bernoulli(q) random variables. Further assume the A_j and B_k variables collectively are mutually independent. The variable X may be viewed as the number of the first A_j that produces a result of 1 and the variable Y may be viewed similarly for the B_k sequence.

(a) Let C_j be a random variable that is 1 if either $A_j = 1$ or $B_j = 1$ (or both), and is equal to 0 otherwise. Prove that $C_j \sim \text{Bernoulli}(r)$ for some quantity r and calculate the value of r in terms of p and q.

(b) Explain why the sequence $C_1, C_2, \ldots$ are mutually independent random variables.

(c) Let Z be the random variable that equals the number of the first C_j that results in a 1 and explain why Z is the smaller of X and Y.

(d) Use (c) to conclude that $Z \sim \text{Geometric}(r)$ for the value of r calculated in part (a).

Ex. 3.3.16. Each day during the hatching season along the Odisha and Northern Tamil Nadu coast line a Poisson (λ) number of turtle eggs hatch giving birth to young turtles. As these turtles swim into the sea the probability that they will survive each day is p. Assume that number of hatchings on each day and the life of the turtles born are all independent. Let $X_1 = 0$ and for $i \geq 2$, X_i be the total number of turtles alive at sea on the i^{th} morning of the hatching season before the hatchings on the i-th day. Find the distribution of X_n.