Discrete Random Variables and their distribution

From Michael Barron Book

3.1.1 Main concepts

DEFINITION 3.1

A **random variable** is a function of an outcome,

$$X = f(\omega).$$

In other words, it is a quantity that depends on chance.

The domain of a random variable is the sample space Ω. Its range can be the set of all real numbers $\mathbb{R}$, or only the positive numbers $(0, +\infty)$, or the integers $\mathbb{Z}$, or the interval $(0, 1)$, etc., depending on what possible values the random variable can potentially take.

Once an experiment is completed, and the outcome ω is known, the value of random variable $X(\omega)$ becomes determined.

Example 3.1. Consider an experiment of tossing 3 fair coins and counting the number of heads. Certainly, the same model suits the number of girls in a family with 3 children, the number of 1's in a random binary string of 3 characters, etc.

Let X be the number of heads (girls, 1's). Prior to an experiment, its value is not known. All we can say is that X has to be an integer between 0 and 3. Since assuming each value is an event, we can compute probabilities,

$$\begin{aligned}
\boldsymbol{P}\{X=0\} &= \boldsymbol{P}\{\text{three tails}\} = \boldsymbol{P}\{TTT\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8} \\
\boldsymbol{P}\{X=1\} &= \boldsymbol{P}\{HTT\} + \boldsymbol{P}\{THT\} + \boldsymbol{P}\{TTH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=2\} &= \boldsymbol{P}\{HHT\} + \boldsymbol{P}\{HTH\} + \boldsymbol{P}\{THH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=3\} &= \boldsymbol{P}\{HHH\} = \frac{1}{8}
\end{aligned}$$

Summarizing,

x	$\boldsymbol{P}\{X=x\}$
0	1/8
1	3/8
2	3/8
3	1/8
Total	1

$\Diamond$

This table contains everything that is known about random variable X prior to the experiment. Before we know the outcome ω, we cannot tell what X equals to. However, we can list all the possible values of X and determine the corresponding probabilities.

DEFINITION 3.2

Collection of all the probabilities related to X is the **distribution** of X. The function

$$P(x) = \boldsymbol{P}\{X = x\}$$

is the **probability mass function**, or **pmf**. The **cumulative distribution function**, or **cdf** is defined as

$$F(x) = \boldsymbol{P}\{X \leq x\} = \sum_{y \leq x} \boldsymbol{P}(y). \tag{3.1}$$

The set of possible values of X is called the **support** of the distribution F.

For every outcome ω, the variable X takes one and only one value x. This makes events $\{X = x\}$ disjoint and exhaustive, and therefore,

$$\sum_x P(x) = \sum_x \boldsymbol{P}\{X = x\} = 1.$$

Looking at (3.1), we can conclude that the cdf $F(x)$ is a non-decreasing function of x, always between 0 and 1, with

$$\lim_{x \downarrow -\infty} F(x) = 0 \text{ and } \lim_{x \uparrow +\infty} F(x) = 1.$$

Between any two subsequent values of X, $F(x)$ is constant. It jumps by $P(x)$ at each possible value x of X (see Figure 3.1, right).

Recall that one way to compute the probability of an event is to add probabilities of all the outcomes in it. Hence, for any set A,

$$\boldsymbol{P}\{X \in A\} = \sum_{x \in A} P(x).$$

When A is an interval, its probability can be computed directly from the cdf $F(x)$,

$$\boldsymbol{P}\{a < X \le b\} = F(b) - F(a).$$

Example 3.2. The pmf and cdf of X in Example 3.1 are shown in Figure 3.1.

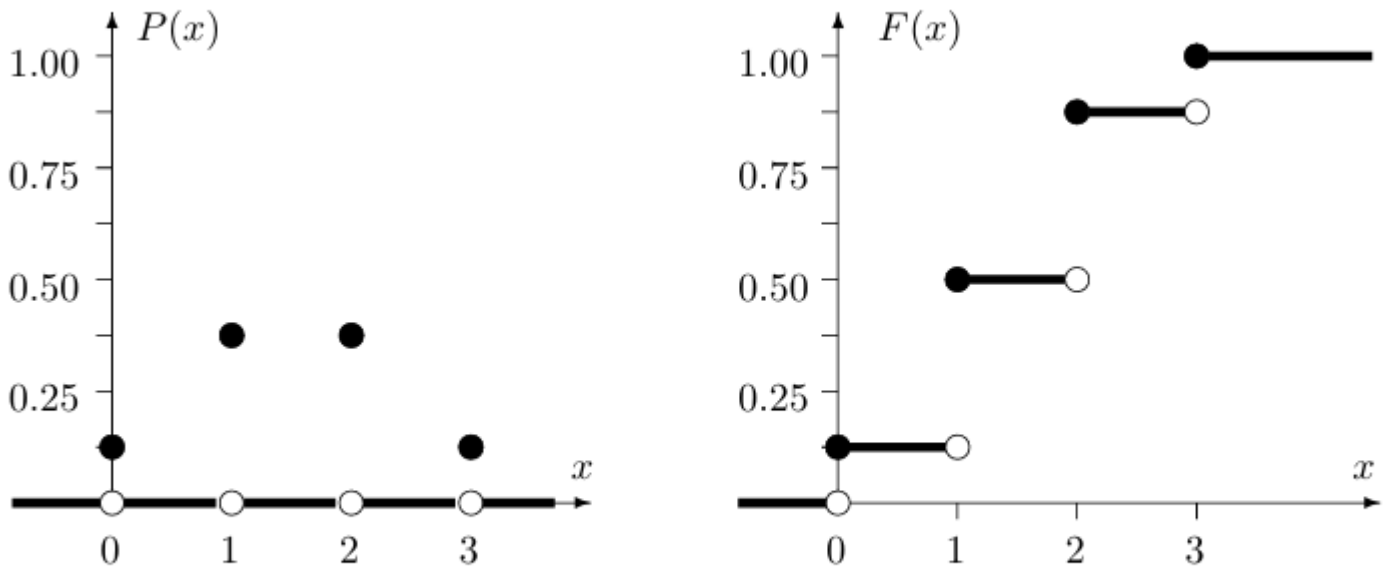

FIGURE 3.1: *The probability mass function $P(x)$ and the cumulative distribution function $F(x)$ for Example 3.1. White circles denote excluded points.*

Example 3.3 (Errors in independent modules). A program consists of two modules. The number of errors X_1 in the first module has the pmf $P_1(x)$, and the number of errors X_2 in the second module has the pmf $P_2(x)$, independently of X_1, where

x	$P_1(x)$	$P_2(x)$
0	0.5	0.7
1	0.3	0.2
2	0.1	0.1
3	0.1	0

Find the pmf and cdf of $Y = X_1 + X_2$, the total number of errors.

Solution. We break the problem into steps. First, determine all possible values of Y, then compute the probability of each value. Clearly, the number of errors Y is an integer that can be as low as $0+0=0$ and as high as $3+2=5$. Since $P_2(3)=0$, the second module has at most 2 errors. Next,

$$\begin{aligned}
P_Y(0) &= P\{Y=0\} &&= \boldsymbol{P}\{X_1=X_2=0\} = P_1(0)P_2(0)\\
&&&= (0.5)(0.7)=0.35\\
P_Y(1) &= P\{Y=1\} &&= P_1(0)P_2(1)+P_1(1)P_2(0)\\
&&&= (0.5)(0.2)+(0.3)(0.7)=0.31\\
P_Y(2) &= P\{Y=2\} &&= P_1(0)P_2(2)+P_1(1)P_2(1)+P_1(2)P_2(0)\\
&&&= (0.5)(0.1)+(0.3)(0.2)+(0.1)(0.7)=0.18\\
P_Y(3) &= P\{Y=3\} &&= P_1(1)P_2(2)+P_1(2)P_2(1)+P_1(3)P_2(0)\\
&&&= (0.3)(0.1)+(0.1)(0.2)+(0.1)(0.7)=0.12\\
P_Y(4) &= P\{Y=4\} &&= P_1(2)P_2(2)+P_1(3)P_2(1)\\
&&&= (0.1)(0.1)+(0.1)(0.2)=0.03\\
P_Y(5) &= P\{Y=5\} &&= P_1(3)P_2(2)=(0.1)(0.1)=0.01
\end{aligned}$$

Now check:

$$\sum_{y=0}^{5} P_Y(y) = 0.35+0.31+0.18+0.12+0.03+0.01=1,$$

thus we *probably* counted all the possibilities and did not miss any. (We just wanted to emphasize that simply getting $\sum P(x) = 1$ does not guarantee that we made no mistake in our solution. However, if this equality is not satisfied, we have a mistake for sure.)

The cumulative function can be computed as

$$\begin{aligned}
F_Y(0) &= P_Y(0) = 0.35 \\
F_Y(1) &= F_Y(0) + P_Y(1) = 0.35 + 0.31 = 0.66 \\
F_Y(2) &= F_Y(1) + P_Y(2) = 0.66 + 0.18 = 0.84 \\
F_Y(3) &= F_Y(2) + P_Y(3) = 0.84 + 0.12 = 0.96 \\
F_Y(4) &= F_Y(3) + P_Y(4) = 0.96 + 0.03 = 0.99 \\
F_Y(5) &= F_Y(4) + P_Y(5) = 0.99 + 0.01 = 1.00
\end{aligned}$$

Between the values of Y, $F(x)$ is constant. $\Diamond$

3.1.2 Types of random variables

So far, we are dealing with *discrete random variables*. These are variables whose range is finite or countable. In particular, it means that their values can be listed, or arranged in a sequence. Examples include the number of jobs submitted to a printer, the number of errors, the number of error-free modules, the number of failed components, and so on. Discrete variables don't have to be integers. For example, the *proportion* of defective components in a lot of 100 can be 0, 1/100, 2/100, ..., 99/100, or 1. This variable assumes 101 different values, so it is discrete, although not an integer.

On the contrary, *continuous random variables* assume a whole interval of values. This could be a bounded interval (a, b), or an unbounded interval such as $(a, +\infty)$, $(-\infty, b)$, or $(-\infty, +\infty)$. Sometimes, it may be a union of several such intervals. Intervals are uncountable, therefore, all values of a random variable cannot be listed in this case. Examples of continuous variables include various times (software installation time, code execution time, connection time, waiting time, lifetime), also physical variables like weight, height, voltage, temperature, distance, the number of miles per gallon, etc. We shall discuss continuous random variables in detail in Chapter 4.

3.2 Distribution of a random vector

Often we deal with several random variables simultaneously. We may look at the size of a RAM and the speed of a CPU, the price of a computer and its capacity, temperature and humidity, technical and artistic performance, etc.

DEFINITION 3.3

If X and Y are random variables, then the pair (X, Y) is a **random vector**. Its distribution is called the **joint distribution** of X and Y. Individual distributions of X and Y are then called the **marginal distributions**.

Although we talk about two random variables in this section, all the concepts extend to a vector $(X_1, X_2, \ldots, X_n)$ of n components and its joint distribution.

Similarly to a single variable, the *joint distribution* of a vector is a collection of probabilities for a vector (X, Y) to take a value (x, y). Recall that two vectors are equal,

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

if $X = x$ and $Y = y$. This "and" means the intersection, therefore, the *joint probability mass function* of X and Y is

$$P(x, y) = \boldsymbol{P}\{(X, Y) = (x, y)\} = \boldsymbol{P}\{X = x \cap Y = y\}.$$

Again, $\{(X, Y) = (x, y)\}$ are exhaustive and mutually exclusive events for different pairs (x, y), therefore,

$$\sum_x \sum_y P(x, y) = 1.$$

3.2.2 Independence of random variables

DEFINITION 3.4

Random variables X and Y are **independent** if

$$P_{(X,Y)}(x, y) = P_X(x)P_Y(y)$$

for *all* values of x and y. This means, events $\{X = x\}$ and $\{Y = y\}$ are independent for all x and y; in other words, variables X and Y take their values independently of each other.

In problems, to show independence of X and Y, we have to check whether the joint pmf factors into the product of marginal pmfs for *all* pairs x and y. To prove dependence, we only need to present one counterexample, a pair (x, y) with $P(x, y) \neq P_X(x)P_Y(y)$.

Example 3.6. A program consists of two modules. The number of errors, X, in the first module and the number of errors, Y, in the second module have the joint distribution, $P(0,0) = P(0,1) = P(1,0) = 0.2$, $P(1,1) = P(1,2) = P(1,3) = 0.1$, $P(0,2) = P(0,3) = 0.05$. Find (a) the marginal distributions of X and Y, (b) the probability of no errors in the first module, and (c) the distribution of the total number of errors in the program. Also, (d) find out if errors in the two modules occur independently.

Solution. It is convenient to organize the joint pmf of X and Y in a table. Adding rowwise and columnwise, we get the marginal pmfs,

$P_{(X,Y)}(x,y)$		y				
		0	1	2	3	$P_X(x)$
x	0	0.20	0.20	0.05	0.05	0.50
	1	0.20	0.10	0.10	0.10	0.50
$P_Y(y)$		0.40	0.30	0.15	0.15	1.00

This solves (a).

(b) $P_X(0) = 0.50$.

(c) Let $Z = X + Y$ be the total number of errors. To find the distribution of Z, we first identify its possible values, then find the probability of each value. We see that Z can be as small as 0 and as large as 4. Then,

$$\begin{aligned}
P_Z(0) &= \boldsymbol{P}\{X+Y=0\} = \boldsymbol{P}\{X=0 \cap Y=0\} = P(0,0) = 0.20, \\
P_Z(1) &= \boldsymbol{P}\{X=0 \cap Y=1\} + \boldsymbol{P}\{X=1 \cap Y=0\} \\
&= P(0,1) + P(1,0) = 0.20 + 0.20 = 0.40, \\
P_Z(2) &= P(0,2) + P(1,1) = 0.05 + 0.10 = 0.15, \\
P_Z(3) &= P(0,3) + P(1,2) = 0.05 + 0.10 = 0.15, \\
P_Z(4) &= P(1,3) = 0.10.
\end{aligned}$$

It is a good check to verify that $\sum_z P_Z(z) = 1$.

(d) To decide on the independence of X and Y, check if their joint pmf factors into a product of marginal pmfs. We see that $P_{(X,Y)}(0,0) = 0.2$ indeed equals $P_X(0)P_Y(0) = (0.5)(0.4)$. Keep checking... Next, $P_{(X,Y)}(0,1) = 0.2$ whereas $P_X(0)P_Y(1) = (0.5)(0.3) = 0.15$. There is no need to check further. We found a pair of x and y that violates the formula for independent random variables. Therefore, the numbers of errors in two modules are dependent. $\Diamond$

Example 3.14: Two ballpoint pens are selected at random from a box that contains 3 blue pens, 2 red pens, and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find

(a) the joint probability function $f(x, y)$,

(b) $P[(X, Y) \in A]$, where A is the region $\{(x, y)|x + y \leq 1\}$.

Note: Above example is from Walpole Book (WP)

***Solution*:** The possible pairs of values (x, y) are $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$, $(0, 2)$, and $(2, 0)$.

(a) Now, $f(0, 1)$, for example, represents the probability that a red and a green pens are selected. The total number of equally likely ways of selecting any 2 pens from the 8 is $\binom{8}{2} = 28$. The number of ways of selecting 1 red from 2 red pens and 1 green from 3 green pens is $\binom{2}{1}\binom{3}{1} = 6$. Hence, $f(0, 1) = 6/28 = 3/14$. Similar calculations yield the probabilities for the other cases, which are presented in Table 3.1. Note that the probabilities sum to 1. In Chapter

5, it will become clear that the joint probability distribution of Table 3.1 can be represented by the formula

$$f(x,y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{\binom{8}{2}},$$

for $x = 0, 1, 2$; $y = 0, 1, 2$; and $0 \le x + y \le 2$.

(b) The probability that (X, Y) fall in the region A is

$$\begin{aligned} P[(X,Y) \in A] &= P(X+Y \le 1) = f(0,0) + f(0,1) + f(1,0) \\ &= \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}. \end{aligned}$$

Table 3.1: Joint Probability Distribution for Example 3.14

	$f(x,y)$	x			Row
		0	1	2	Totals
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Column Totals		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Exercises

3.1. A computer virus is trying to corrupt two files. The first file will be corrupted with probability 0.4. Independently of it, the second file will be corrupted with probability 0.3.

(a) Compute the probability mass function (pmf) of X, the number of corrupted files.

(b) Draw a graph of its cumulative distribution function (cdf).

3.2. Every day, the number of network blackouts has a distribution (probability mass function)

x	0	1	2
$P(x)$	0.7	0.2	0.1

Note: Do only parts related to topic of this slides

A small internet trading company estimates that each network blackout results in a \$500 loss. Compute expectation and variance of this company's daily loss due to blackouts.

3.3. There is one error in one of five blocks of a program. To find the error, we test three randomly selected blocks. Let X be the number of errors in these three blocks. Compute $\mathbf{E}(X)$ and $\mathrm{Var}(X)$.

3.4. Tossing a fair die is an experiment that can result in any integer number from 1 to 6 with equal probabilities. Let X be the number of dots on the top face of a die. Compute $\mathbf{E}(X)$ and $\mathrm{Var}(X)$.

3.10. Every day, the number of traffic accidents has the probability mass function

x	0	1	2	more than 2
$P(x)$	0.6	0.2	0.2	0

independently of other days. What is the probability that there are more accidents on Friday than on Thursday?

3.11. Two dice are tossed. Let X be *the smaller* number of points. Let Y be *the larger* number of points. If both dice show the same number, say, z points, then $X = Y = z$.

Note: Do only parts related to topic of this slides

(a) Find the joint probability mass function of (X, Y).

(b) Are X and Y independent? Explain.

(c) Find the probability mass function of X.

(d) If $X = 2$, what is the probability that $Y = 5$?

3.12. Two random variables, X and Y, have the joint distribution $P(x, y)$,

$P(x,y)$		x	
		0	1
y	0	0.5	0.2
	1	0.2	0.1

(a) Are X and Y independent? Explain.

(b) Are $(X + Y)$ and $(X - Y)$ independent? Explain.

3.14. An internet service provider charges its customers for the time of the internet use rounding it up to the nearest hour. The joint distribution of the used time (X, hours) and the charge per hour (Y, cents) is given in the table below.

$P(x,y)$		x			
		1	2	3	4
y	1	0	0.06	0.06	0.10
	2	0.10	0.10	0.04	0.04
	3	0.40	0.10	0	0

Each customer is charged $Z = X \cdot Y$ cents, which is the number of hours multiplied by the price of each hour. Find the distribution of Z.

3.15. Let X and Y be the number of hardware failures in two computer labs in a given month. The joint distribution of X and Y is given in the table below.

$P(x,y)$		x		
		0	1	2
y	0	0.52	0.20	0.04
	1	0.14	0.02	0.01
	2	0.06	0.01	0

(a) Compute the probability of at least one hardware failure.

(b) From the given distribution, are X and Y independent? Why or why not?