3.1 Random variables

Unit 3: Discrete Random Variables

Adapted from Blitzstein-Hwang Chapter 3.

Random variables are an incredibly useful concept that simplifies notation and expands our ability to quantify uncertainty and summarize the results of experiments. Random variables are essential throughout statistics, so it is crucial to think through what they mean, both intuitively and mathematically.

Sometimes a definition of "random variable" (r.v.) is given that is a barely paraphrased version of "a random variable is a variable that takes on random values", but this fails to say where the randomness come from. To make the notion of random variable precise, we define it as a *function* mapping the sample space to the real line.

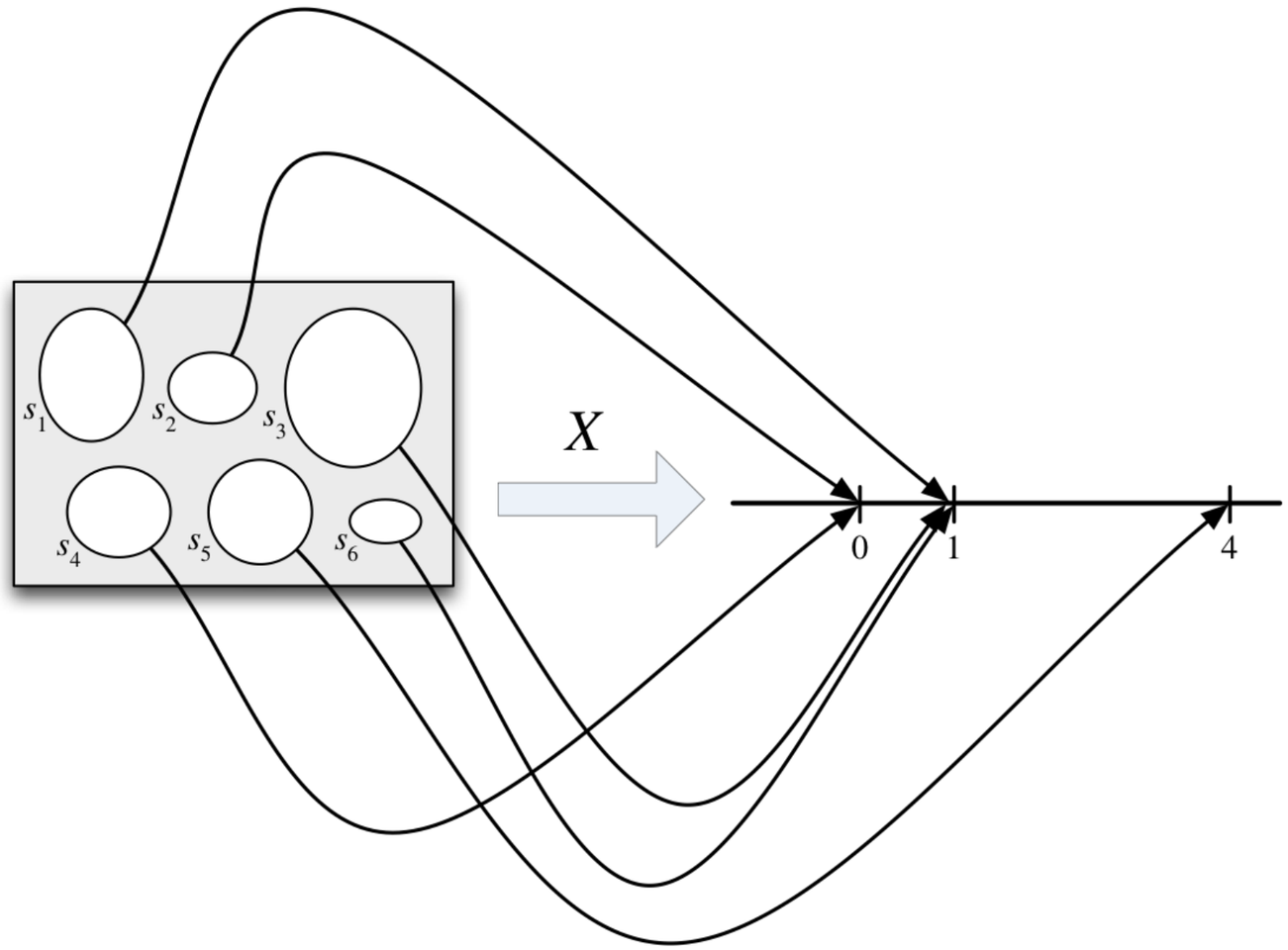

Figure 3.1.1: A random variable maps the sample space into the real line. The r.v. X depicted here is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The randomness comes from choosing a random pebble.

View Larger Image

Image Description

Definition 3.1.2 (Random variable).

Given an experiment with sample space S, a *random variable* (r.v.) is a function from the sample space S to the real numbers $\mathbb{R}$. It is common, but not required, to denote random variables by capital letters.

Thus, a random variable X assigns a numerical value $X(s)$ to each possible outcome s of the experiment. The randomness comes from the fact that we have a random experiment (with probabilities described by the probability function P); the mapping itself is deterministic. The same r.v. is shown in a simpler way in the left panel of Figure 3.1.4, in which we inscribe the values inside the pebbles.

This definition is abstract but fundamental; one of the most important skills to develop when studying probability and statistics is the ability to go back and forth between abstract ideas and concrete examples. Relatedly, it is important to work on recognizing the essential pattern or structure of a problem and how it connects to problems you have studied previously. We will often discuss stories that involve tossing coins or drawing balls from urns because they are simple, convenient scenarios to work with, but many other problems are *isomorphic*: they have the same essential structure, but in a different guise.

To start, let's consider a coin-tossing example. The structure of the problem is that we have a sequence of trials where there are two possible outcomes for each trial. Here we think of the possible outcomes as H (Heads) and T (Tails), but we could just as well think of them as "success" and "failure" or as 1 and 0, for example.

Example 3.1.3 (Coin tosses).

Consider an experiment where we toss a fair coin twice. The sample space consists of four possible outcomes: $S = \{HH, HT, TH, TT\}$. Here are some random variables on this space (for practice, you can think up some of your own). Each r.v. is a numerical summary of some aspect of the experiment.

- Let X be the number of Heads. This is a random variable with possible values 0, 1, and 2. Viewed as a function, X assigns the value 2 to the outcome HH, 1 to the outcomes HT and TH, and 0 to the outcome TT. That is,
$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0.$$
- Let Y be the number of Tails. In terms of X, we have $Y = 2 - X$. In other words, Y and $2 - X$ are the same r.v.: $Y(s) = 2 - X(s)$ for all s.
- Let I be 1 if the first toss lands Heads and 0 otherwise. Then I assigns the value 1 to the outcomes HH and HT and 0 to the outcomes TH and TT. This r.v. is an example of what is called an *indicator random variable* since it indicates whether the first toss lands Heads, using 1 to mean "yes" and 0 to mean "no".

We can also encode the sample space as $\{(1,1), (1,0), (0,1), (0,0)\}$, where 1 is the code for Heads and 0 is the code for Tails. Then we can give explicit formulas for X, Y, I:

$$X(s_1, s_2) = s_1 + s_2, Y(s_1, s_2) = 2 - s_1 - s_2, I(s_1, s_2) = s_1,$$

where for simplicity we write $X(s_1, s_2)$ to mean $X((s_1, s_2))$, etc. For most r.v.s we will consider, it is tedious or infeasible to write down an explicit formula in this way. Fortunately, it is usually unnecessary to do so. As before, for a sample space with a finite number of outcomes we can visualize the outcomes as pebbles, with the mass of a pebble corresponding to its probability, such that the total mass of the pebbles is 1. A random variable simply labels each pebble with a number. Figure 3.1.4 shows two random variables defined on the same sample space: the pebbles or outcomes are the same, but the real numbers assigned to the outcomes are different.

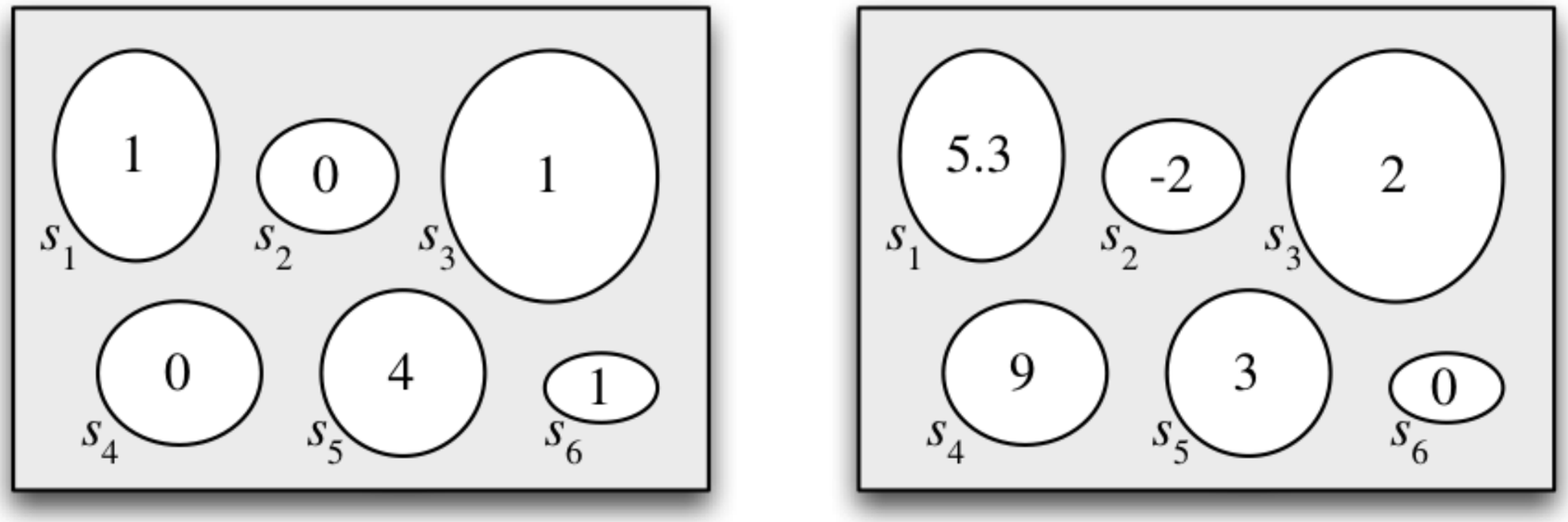

Figure 3.1.4: Two random variables defined on the same sample space.

View Larger Image
Image Description

Before we perform the experiment, the outcome s has not yet been realized, so we don't know the value of X, though we could calculate the probability that X will take on a given value or range of values. After we perform the experiment and the outcome s has been realized, the random variable crystallizes into the numerical value $X(s)$. In this way, random variables provide *numerical* summaries of the experiment in question.