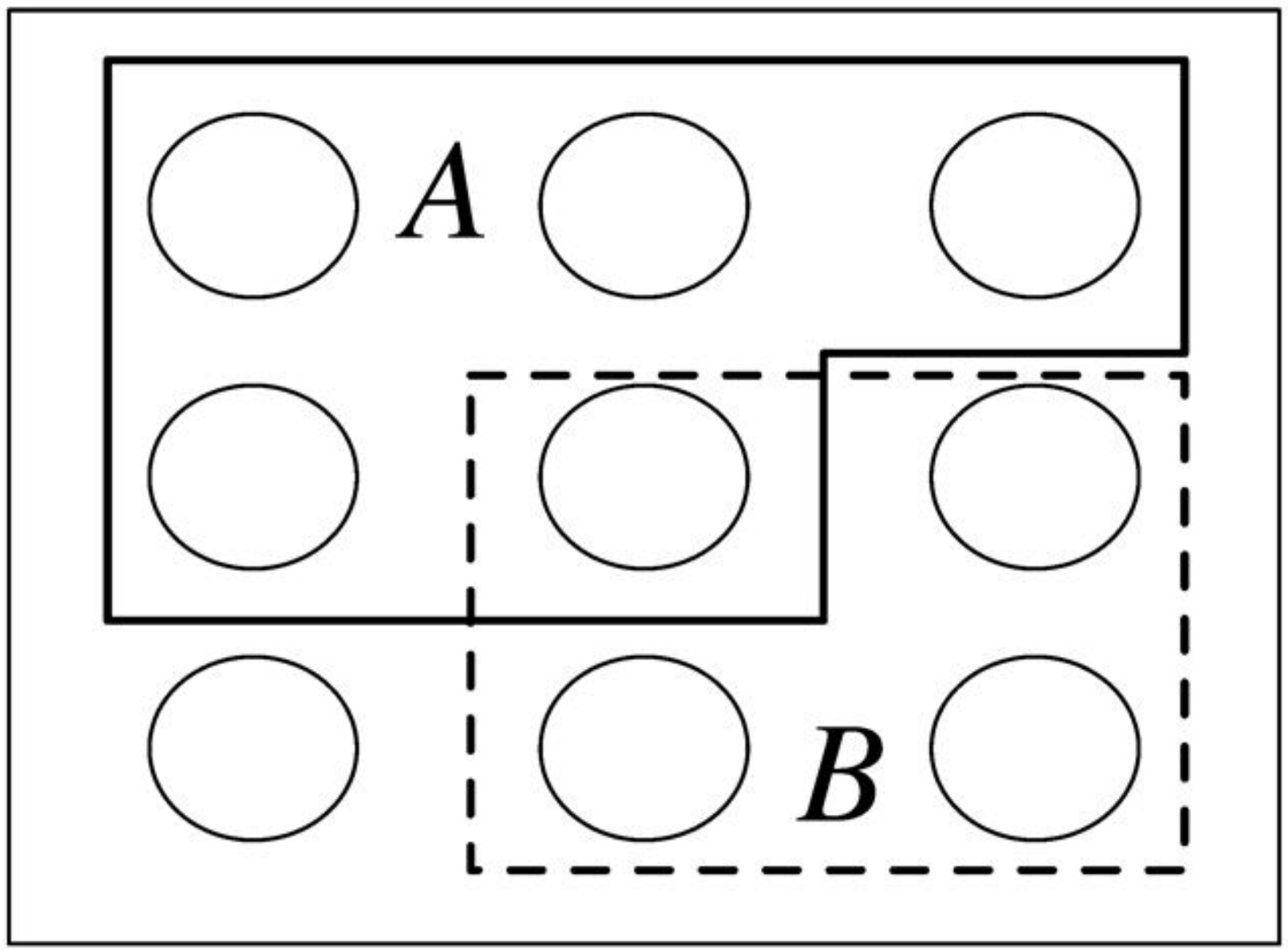
Experiments, Samples, Events, and Sets

The *sample space* S of an experiment is the set of all possible outcomes of the experiment. An *event* A is a subset of the sample space S, and we say that A *occurred* if the actual outcome is in A.

FIGURE 1.1 A sample space as Pebble World, with two events A and B spotlighted.



Set Notation

English	Sets
Events and occurrences	
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
something must happen	$s_{\text{actual}} \in S$
New events from old events	
A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \cdots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \cdots, A_n$	$A_1 \cap \cdots \cap A_n$
Relationships between events	
A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \phi$
$A_1, \cdots, A_n$ are a partition of S	$A_1 \cup \cdots \cup A_n = S, A_i \cap A_j = \phi$ for $i \neq j$

What is probability?

Definition 1.6.1 (General definition of probability).

A *probability space* consists of a sample space S and a *probability function* P which takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output. The function P must satisfy the following axioms:

1. $P(\varphi) = 0, P(S) = 1$.
2. If $A_1, A_2, \ldots$ are disjoint events, then

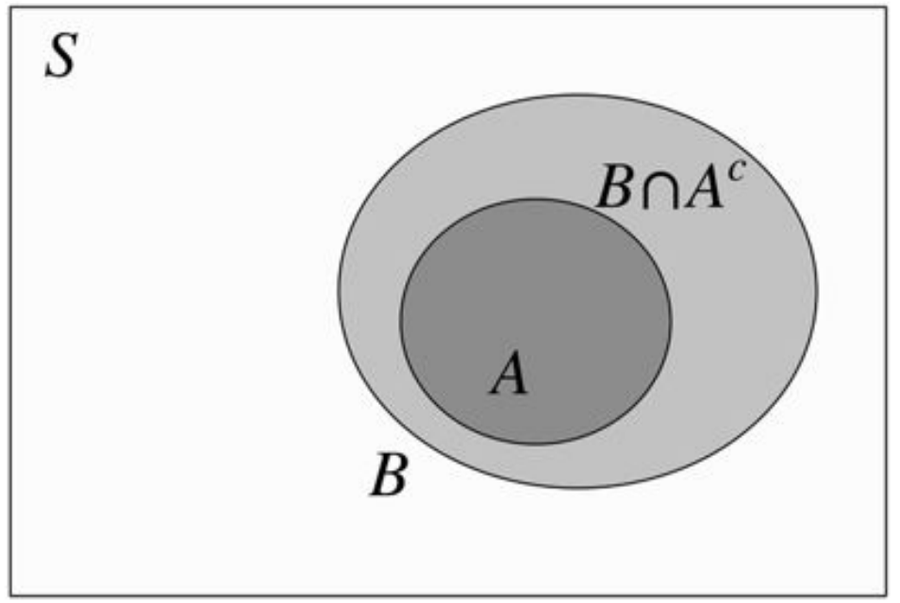
$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j).$$

Any function P (mapping events to numbers in the interval [0,1]) that satisfies the two axioms is considered a valid probability function. However, the axioms don't tell us how probability should be *interpreted*; different schools of thought exist.

The *frequentist* view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment: if we say a coin has probability 1/2 of Heads, that means the coin would land Heads 50% of the time if we tossed it over and over and over.

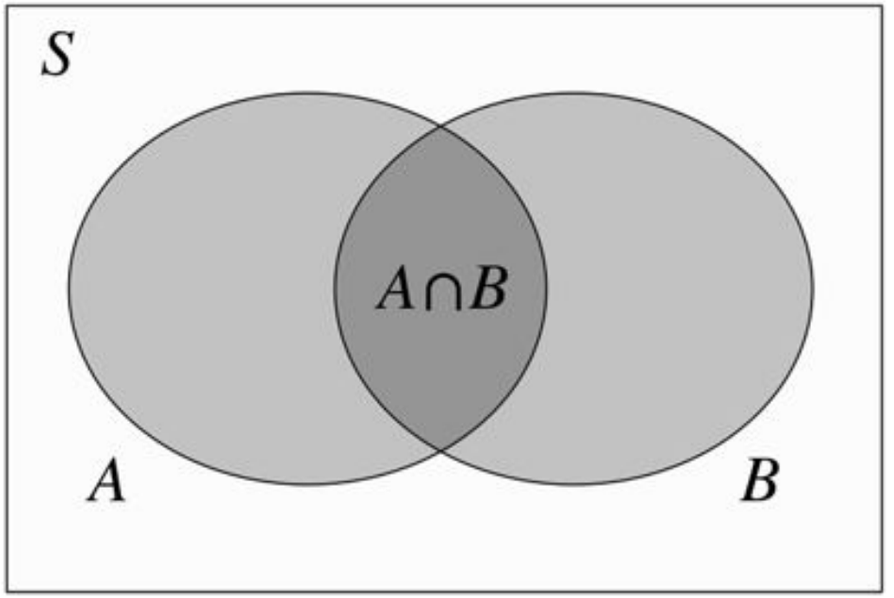
The *Bayesian* view of probability is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses like “candidate A will win the election” or “the defendant is guilty” even if it isn't possible to repeat the same election or the same crime over and over again.

The Bayesian and frequentist perspectives are complementary, and both will be helpful for developing intuition in later chapters. Regardless of how we choose to interpret probability, we can use the two axioms to derive other properties of probability, and these results will hold for *any* valid probability function.



Probability has the following properties, for any events A and B.

1. $P(A^C) = 1 - P(A)$.
2. If $A \subseteq B$, then $P(A) \leq P(B)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.



Conditional Probabilities

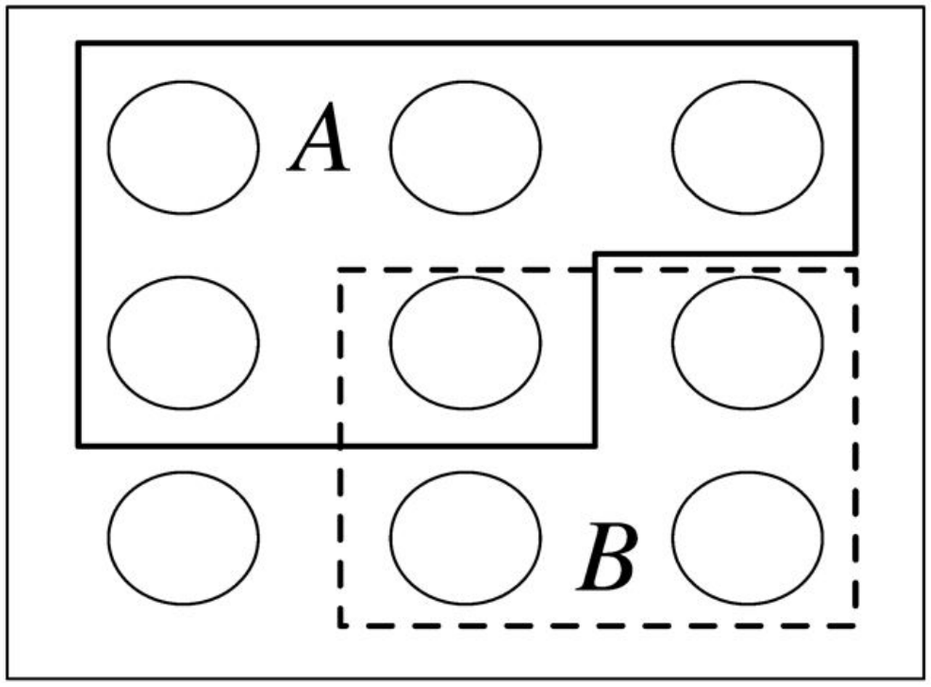
Definition 2.2.1 (Conditional probability).

If A and B are events with $P(B)>0$, then the *conditional probability* of A given B, denoted by $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Here A is the event whose uncertainty we want to update, and B is the evidence we observe (or want to treat as given). We call $P(A)$ the *prior* probability of A and $P(A|B)$ the *posterior* probability of A ("prior" means before updating based on the evidence, and "posterior" means after updating based on the evidence).

It is important to interpret the event appearing after the vertical conditioning bar as the evidence that we have observed or that is being conditioned on: $P(A|B)$ is the probability of A given the evidence B, *not* the probability of some entity called $A|$



Conditional Probabilities

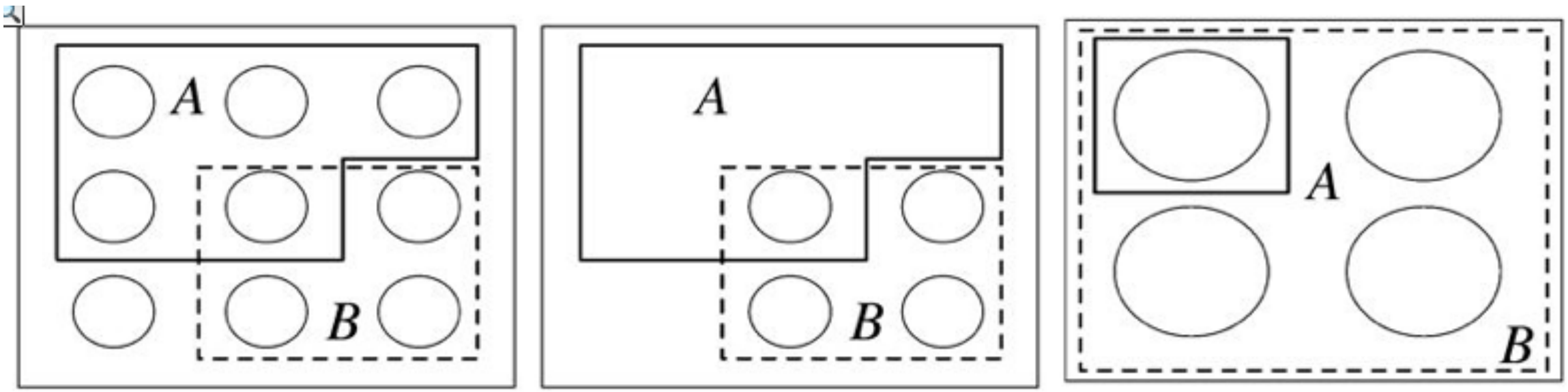
Definition 2.2.1 (Conditional probability).

If A and B are events with $P(B)>0$, then the *conditional probability* of A given B, denoted by $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Here A is the event whose uncertainty we want to update, and B is the evidence we observe (or want to treat as given). We call $P(A)$ the *prior* probability of A and $P(A|B)$ the *posterior* probability of A ("prior" means before updating based on the evidence, and "posterior" means after updating based on the evidence).

It is important to interpret the event appearing after the vertical conditioning bar as the evidence that we have observed or that is being conditioned on: $P(A|B)$ is the probability of A given the evidence B, *not* the probability of some entity called $A|$



Random Variables

FIGURE 3.1 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 according to the probability function P for the sample space.

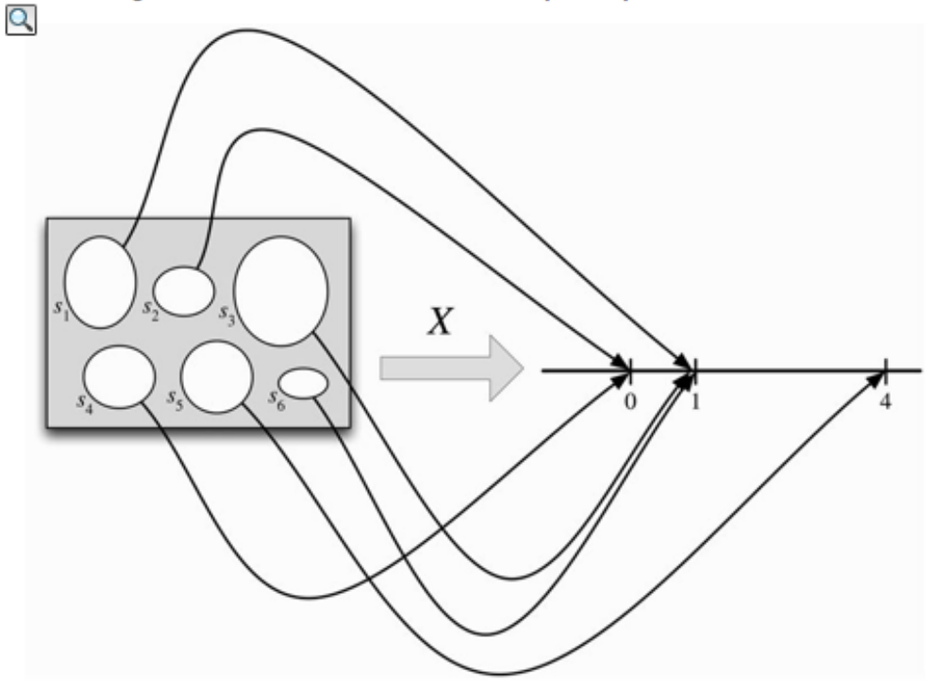


FIGURE 3.2 Two random variables defined on the same sample space.

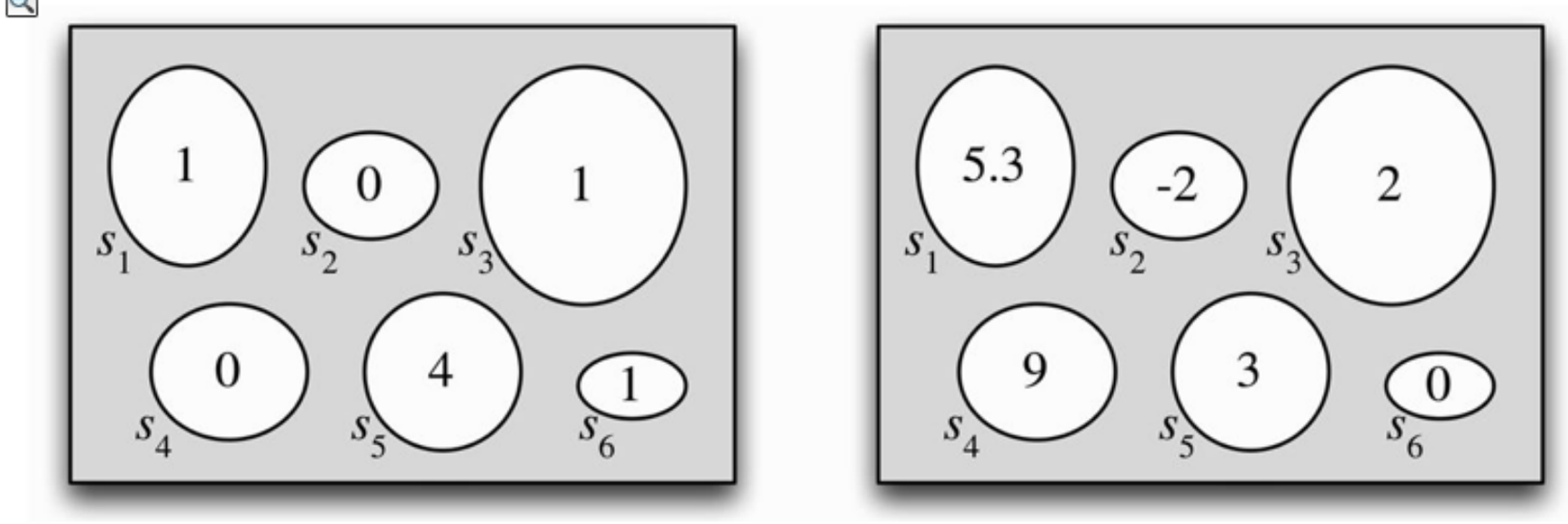
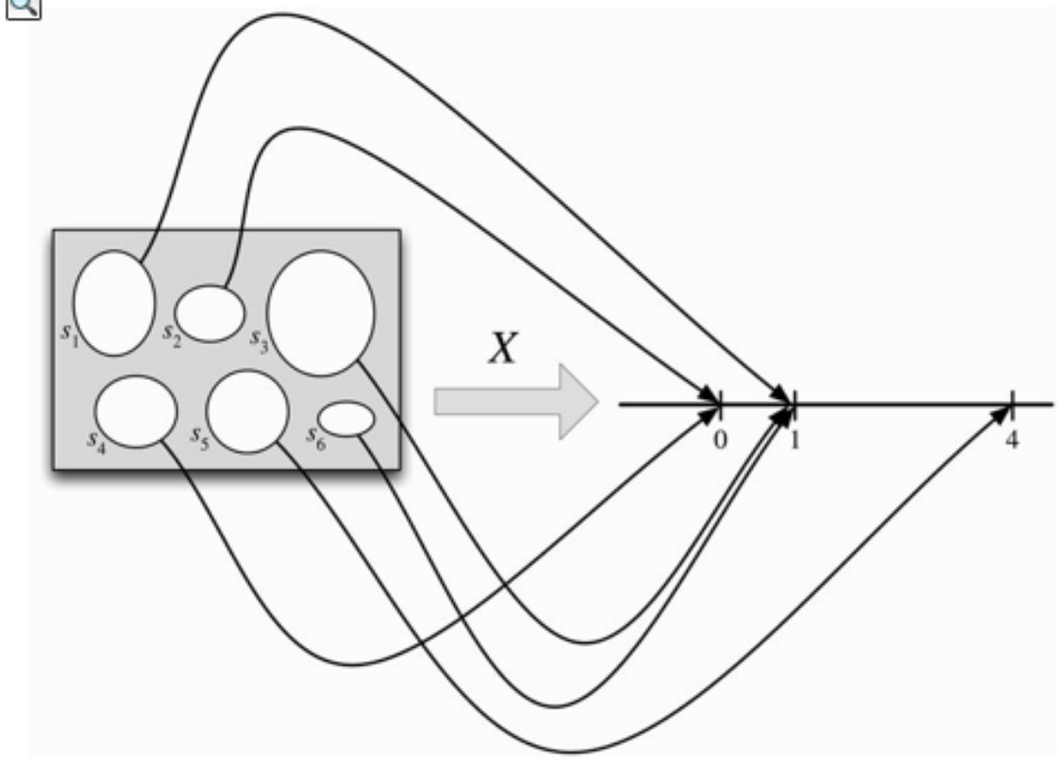
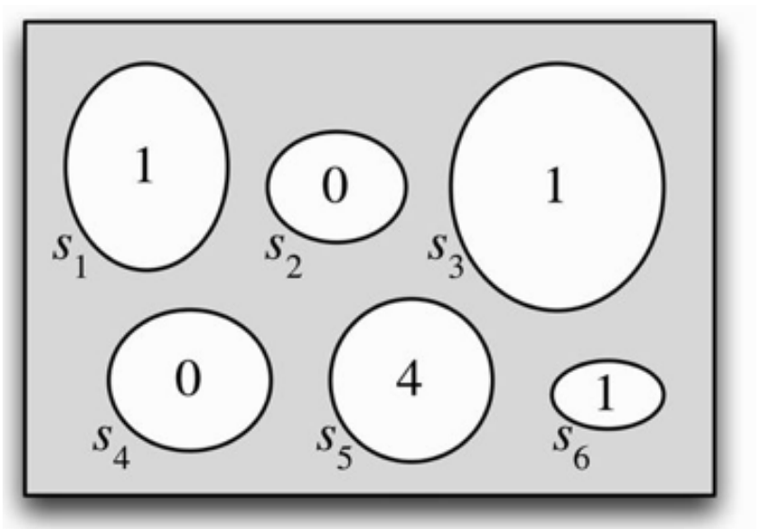


FIGURE 3.1 A random variable maps the sample space into the real line. The r.v. X depicted here is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The randomness comes from choosing a random pebble according to the probability function P for the sample space.



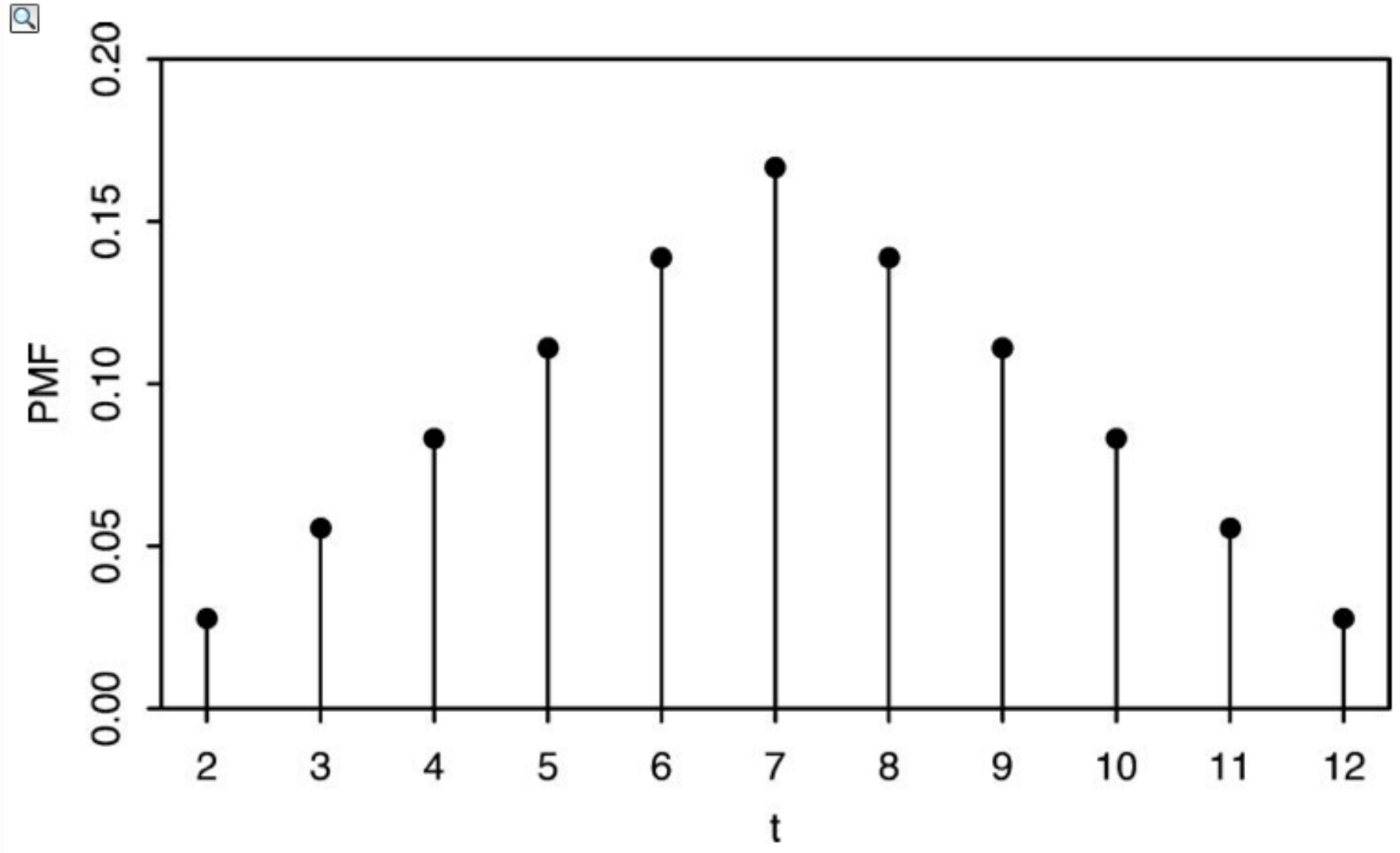
X



$$P(X=1) = P(s_1)+P(s_3)+P(s_6)$$

Probability Mass Function (PMF)

FIGURE 3.4 PMF of the sum of two die rolls.



Bernoulli - The Simplest Discrete Distribution

Definition 3.3.1 (Bernoulli distribution).

An r.v. X is said to have the *Bernoulli distribution* with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 < p < 1$. We write this as $X \sim \text{Bern}(p)$. The symbol $\sim$ is read "is distributed as".

Any r.v. whose possible values are 0 and 1 has a Bern(p) distribution, with p the probability of the r.v. equaling 1. This number p in Bern(p) is called the *parameter* of the distribution; it determines which specific Bernoulli distribution we have. Thus there is not just one Bernoulli distribution, but rather a *family* of Bernoulli distributions, indexed by p. For example, if $X \sim \text{Bern}(1/3)$, it would be correct but incomplete to say "X is Bernoulli"; to fully specify the distribution of X, we should both say its name (Bernoulli) and its parameter value (1/3), which is the point of the notation $X \sim \text{Bern}(1/3)$.

Any event has a Bernoulli r.v. that is naturally associated with it, equal to 1 if the event happens and 0 otherwise. This is called the *indicator random variable* of the event; we will see that such r.v.s are extremely useful.

The Binomial - More than one Bernoulli

Theorem 3.3.5 (Binomial PMF).

If $X \sim \text{Bin}(n, p)$, then the PMF of X is

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

for $k = 0, 1, \ldots, n$ (and $P(X = k) = 0$ otherwise).

☣ **3.3.6.** To save writing, it is often left implicit that a PMF is zero wherever it is not specified to be nonzero, but in any case it is important to understand what the support of a random variable is, and good practice to check that PMFs are valid. If two discrete r.v.s have the same PMF, then they also must have the same support. So we sometimes refer to the support of a discrete *distribution*; this is the support of any r.v. with that distribution.

Proof. An experiment consisting of n independent Bernoulli trials produces a sequence of successes and failures. The probability of any specific sequence of k successes and $n - k$ failures is $p^k(1-p)^{n-k}$. There are ($\binom{n}{k}$) such sequences, since we just need to select where the successes are. Therefore, letting X be the number of successes,

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

for $k = 0, 1, \ldots, n$, and $P(X = k) = 0$ otherwise. This is a valid PMF because it is nonnegative and it sums to 1 by the binomial theorem.

Assignment 2

Definition 1.4.12 (Binomial coefficient).

For any nonnegative integers k and n, the *binomial coefficient* $\binom{n}{k}$, read as "n choose k", is the number of subsets of size k for a set of size n.

For example, $\binom{4}{2} = 6$, as shown in Example 1.4.11. The binomial coefficient $\binom{n}{k}$ is sometimes called a *combination*, but we do not use that terminology here since "combination" is such a useful general-purpose word. Algebraically, binomial coefficients can be computed as follows.

Theorem 1.4.13 (Binomial coefficient formula).

For $k \leq n$, we have

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}.$$

For $k>n$, we have $\binom{n}{k} = 0$.

Proof. Let A be a set with $|A| = n$. Any subset of A has size at most n, so $\binom{n}{k} = 0$ for $k>n$. Now let $k \leq n$. By Theorem 1.4.6, there are $n(n-1)\cdots(n-k+1)$ ways to make an *ordered* choice of k elements without replacement. This overcounts each subset of interest by a factor of $k!$ (since we don't care how these elements are ordered), so we can get the correct count by dividing by $k!$.

In general, how might we want to use probability distributions?

In general, how might we want to use probability distributions?

Predict

Infer

Likelihood

A measure of relative surprise

For a binomial:

$$L(p|D) = P(D|p)$$

$$L(p; D) = P(D|p)$$

$$L(p) = P(D|p)$$

The Likelihood Principle

"In statistics, **the likelihood principle** is that, given a statistical model, all of the evidence in a sample relevant to model parameters is contained in the likelihood function."

https://en.wikipedia.org/wiki/Likelihood_principle

Controversial, although several have offered proofs.

Important Notes about Likelihoods

There are some very important distinctions between likelihoods and probabilities. First, likelihoods do NOT sum (or integrate) to 1. Therefore, the area under a likelihood curve is not meaningful, as it is for probability.

Second, it does not make sense to compare likelihoods across different datasets. For instance, smaller numbers of observations generally produce higher likelihoods, because the total sample space is smaller.