1 Introduction

Random variables are essential throughout statistics.

Definition (Random variable)**.** Given an experiment with sample space S, a random variable 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 of the experiment. The randomness comes from the fact that we have a random experiment (with probabilities described by the probability function $\mathsf{P}(\cdot)$); the mapping itself is deterministic – see Figures 1 and 2. The source of the randomness in a random variable is in the experiment itself, in which a sample outcome $s \in S$ is chosen according to a probability function $\mathsf{P}(\cdot)$. Before we perform the experiment, the outcome s has not yet been realized, so we don't know the value of $X = X(\cdot)$, 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 $X(s)$.

Random variables provide *numerical summaries* of the experiment in question. This is very handy because the sample space of an experiment is often incredibly complicated or high-dimensional, and the outcomes $s \in S$ may be non-numeric. For example, the experiment may be to collect a random sample of people in a certain city and ask them various questions, which may have numeric (e.g., age or height) or non-numeric (e.g., political party or favorite movie) answers. The fact that random variables take on numerical values is a very convenient simplification compared to having to work with the full complexity of S at all times.

Definition (Discrete random variable)**.** A random variable X is said to be *discrete* if there is a finite list of values $a_1, a_2, \ldots, a_n$ or an infinite set of values $a_1, a_2, \ldots$ (i.e., a countable set) such that $\mathsf{P}(X = a_j \text{ for some } j) = 1$. If X is a discrete random variable, then the finite or countably infinite set of values x such that $\mathsf{P}(X = x) > 0$ is called the *support* of X.

The support of a discrete random variable is a set of integers. In contrast, a continuous random variable can take on any real value in an interval (possibly even the entire real line). Given a random variable, we would like to be able to describe its behavior using the language of probability. For example, we might want to answer questions about the probability that the random variable will fall into a given range: if L is the lifetime earnings of a randomly chosen U.S. college graduate, what is the probability that L exceeds a million dollars? The *distribution* of a random variable provides the answers to these questions. For discrete random variables, the most natural way to express their distribution is with a *probability mass function*.

Definition (Probability mass function)**.** The *probability mass function* (PMF) of a discrete random variable X is the function $p_X(\cdot)$ given by $p_X(x) = \mathsf{P}(X = x)$. Note that this function is positive if x is in the support of X, and 0 otherwise.

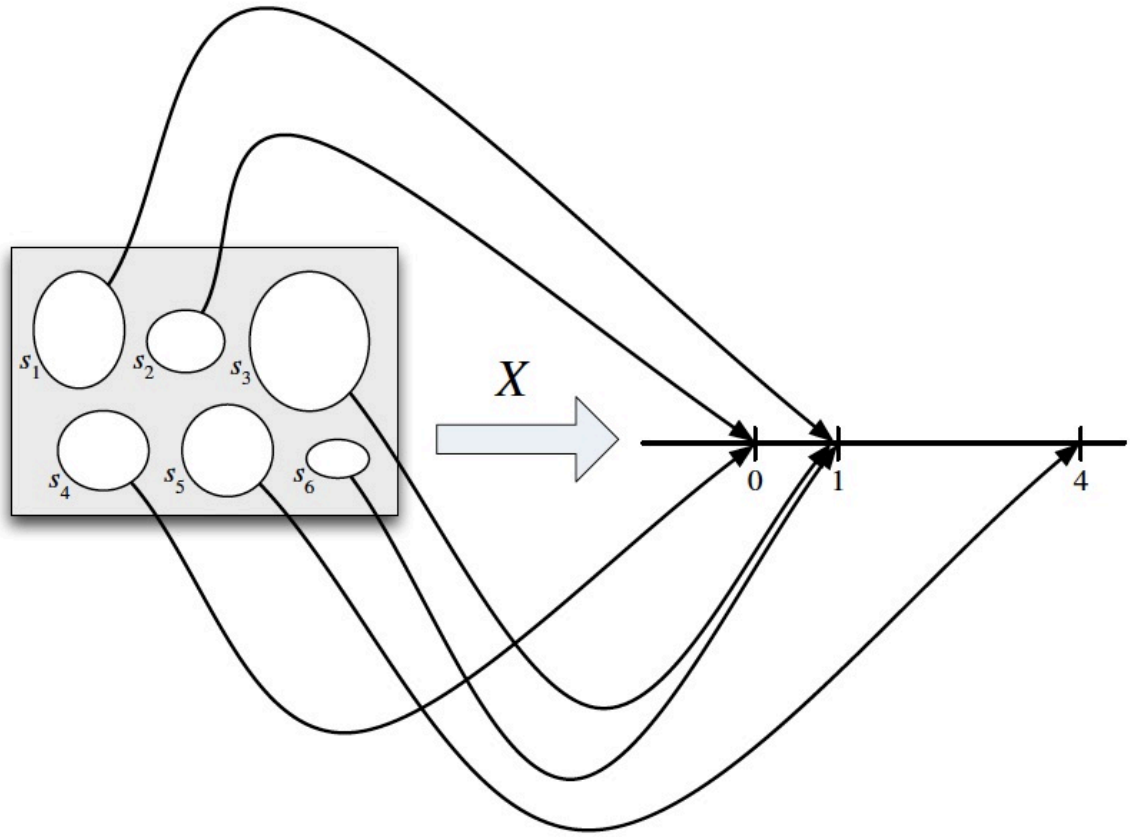

Figure 1: A random variable maps the sample space into the real line. The random variable X depicted here is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The randomness comes from choosing a random pebble according to the probability function $\mathsf{P}(\cdot)$ for the sample space.

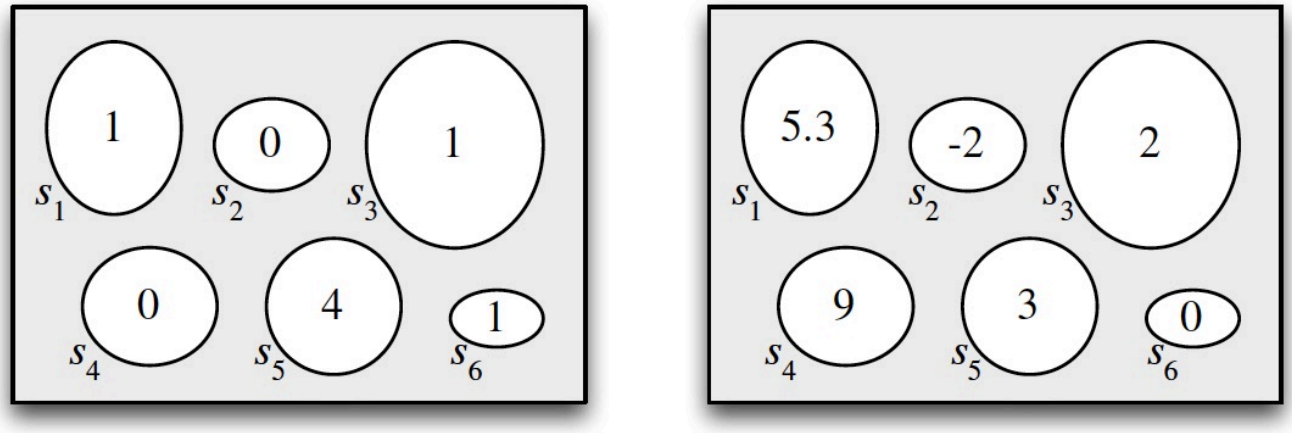

Figure 2: For a sample space with a finite number of outcomes we can visualize the outcomes as pebbles, with the mass of a pebble corresponding to its probability, such that the total mass of the pebbles is 1. A random variable simply labels each pebble with a number. This figure shows two random variables defined on the same sample space: the pebbles or outcomes are the same, but the real numbers assigned to the outcomes are different.

In writing $\mathsf{P}(X = x)$, we are using $X = x$ to denote an *event*, consisting of all outcomes s to which X assigns the number x. This event is also written as $\{X = x\} = \{s \in S : X(s) = x\}$. If $\{X = x\}$ were anything other than an event, it would make no sense to calculate its probability. It does not make sense to write $\mathsf{P}(X)$: we can only take the probability of an event, not of an random variable.

Theorem. *Valid PMFs. Let X be a discrete random variable with support $x_1, x_2, \ldots$. The PMF p_X of X must satisfy:*

- *Nonnegative: $p_X(x) > 0$ if $x = x_j$ for some j, and $p_X(x) = 0$ otherwise.*
- *Sums to 1: $\sum_{j=1}^{\infty} p_X(x_j) = 1$.*

Proof. The first statement is true since probability is nonnegative. The second is true since X must take some value, and the events $\{X = x_j\}$ are disjoint, so

$$\sum_{j=1}^{\infty} \mathsf{P}(X = x_j) = \mathsf{P}\left(\bigcup_{j=1}^{\infty} \{X = x_j\}\right) = \mathsf{P}(X = x_1 \text{ or } X = x_2 \text{ or } \ldots) = 1.$$

□

Knowing the PMF of a discrete random variable determines its distribution.

Example: Sum of die rolls

We roll two fair 6-sided dice. Let $T = X + Y$ be the total of the two rolls, where X and Y are the individual

s	X	Y	$X+Y$
$(1,2)$	1	2	3
$(1,6)$	1	6	7
$(2,5)$	2	5	7
$(3,1)$	3	1	4
$(4,3)$	4	3	7
$(5,4)$	5	4	9
$(6,6)$	6	6	12

Figure 3: Seven of the 36 outcomes in S, along with the corresponding values of X, Y and T.

rolls. The sample space of this experiment has 36 equally likely outcomes:

$$S = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}.$$

After the experiment is performed, we observe values for X and Y, and then the observed value of T is the sum of those values. See Figure 3. Since the dice are fair, the PMF of X is $\mathsf{P}(X = j) = 1/6$ for $j = 1,2,\ldots,6$. We say that X has a *discrete uniform* distribution on $\{1,2,\ldots,6\}$. Similarly, Y is also discrete uniform on $\{1,2,\ldots,6\}$. Note that Y has the same distribution as X, but it is not the same random variable as X. In fact, we have

$$\mathsf{P}(X = Y) = \mathsf{P}(\{s \in S : X(s) = Y(s)\} = 6/36 = 1/6.$$

Two more random variables with the same distribution as X are $7 - X$ and $7 - Y$. To see this, we can use the fact that for a standard die, $7 - X$ is the value on the bottom if X is the value at the top. If the top value is equally likely to be any of the numbers $1,2,\ldots,6$, then so it the bottom value. Note that even though $7 - X$ has the same distribution as X, it is *never* equal to X in a run of the experiment.

Let's now find the PMF of X:

$$\mathsf{P}(T = 2) = \mathsf{P}(T = 12) = \tfrac{1}{36}$$
$$\mathsf{P}(T = 3) = \mathsf{P}(T = 11) = \tfrac{2}{36}$$
$$\mathsf{P}(T = 4) = \mathsf{P}(T = 10) = \tfrac{3}{36}$$
$$\mathsf{P}(T = 5) = \mathsf{P}(T = 9) = \tfrac{4}{36}$$
$$\mathsf{P}(T = 6) = \mathsf{P}(T = 8) = \tfrac{5}{36}$$
$$\mathsf{P}(T = 7) = \tfrac{6}{36}.$$

Once we know the PMF of T, we can calculate the probability that T will fall into a given subset of the real numbers by summing over the appropriate values T can take. Suppose we are interested in the probability that T is in the interval $[1, 4]$:

$$\mathsf{P}(1 \leq T \leq 4) = \mathsf{P}(T = 2) + \mathsf{P}(T = 3) + \mathsf{P}(T = 4) = \frac{6}{36}.$$

2 Bernoulli and Binomial distributions

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

Any event has a Bernoulli random variable that is naturally associated with it: it equals 1 if the event happens and 0 otherwise. This is called the *indicator random variable* of the event. Indicator random variables are extremely useful in statistical theory. We denote the indicator random variable of an event A by I_A or $\mathsf{I}(A)$. Note that $\mathsf{I}_A \sim \mathsf{Bern}(p)$ with $p = \mathsf{P}(A)$.

An experiment that can result in either a "succes" or a "failure" (but not both) is called a Bernoulli trial. A Bernoulli random variable can be thought of as the indicator of success in a Bernoulli trial: it equals 1 if success occurs and 0 if failure occurs in the trial. For this reason, the parameter p is often called the *success probability* of the $\mathsf{Bern}(p)$ distribution.

In the sequel, suppose that n *independent* Bernoulli trials are performed, each with the same success probability p. Let X be the number of successes. The distribution of the random variable X is called the *Binomial distribution* with parameters n and p. We write $X \sim \mathsf{Bin}(n, p)$, where n is integer and $0 < p < 1$.

Theorem. *Binomial PMF If $X \sim \mathsf{Bin}(n, p)$, then the PMF of X is*

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

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

Theorem. *Let $X \sim \mathsf{Bin}(n, p)$, and $q = 1 - p$ (often taken to denote the failure of a Bernoulli trial). Then $n - X \sim \mathsf{Bin}(n, q)$.*

3 The hypergeometric distribution

If we have an urn filled with w white and b black balls, then drawing n balls out of the urn *with replacement* yields a $\mathsf{Bin}(n, w/(w + b))$ distribution for the number of white balls obtained in n trials since the draws are independent Bernoulli trials, each with probability $w/(w + b)$ of success. If we instead sample *without replacement*, then the number of white balls follows a *Hypergeometric distribution* – see Figure 5.

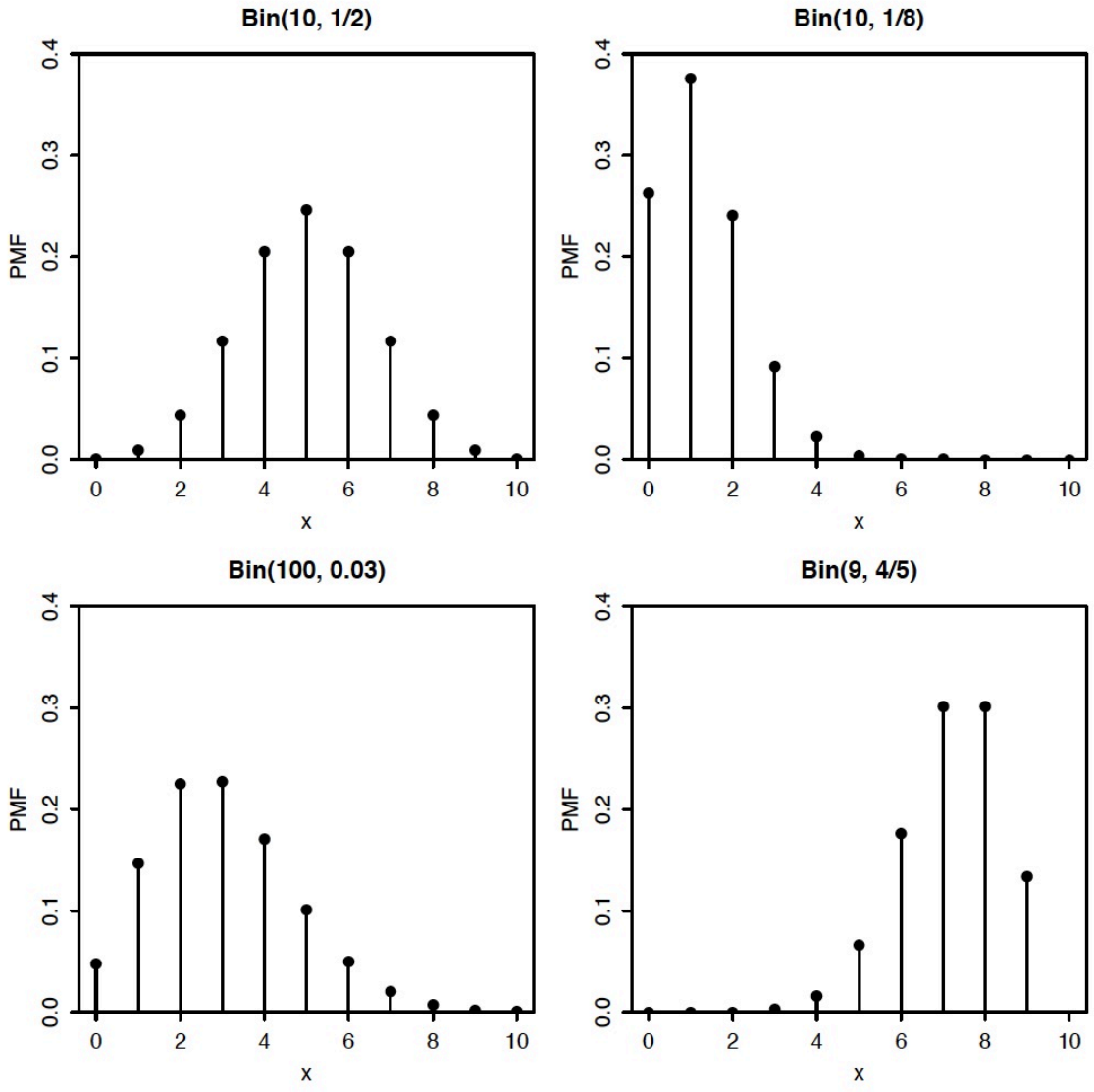

Figure 4: Some Binomial PMFs. In the lower left, we plot the Bin$(100, 0.03)$ PMF between 0 and 10 only, as the probability of more than 10 successes is close to 0.

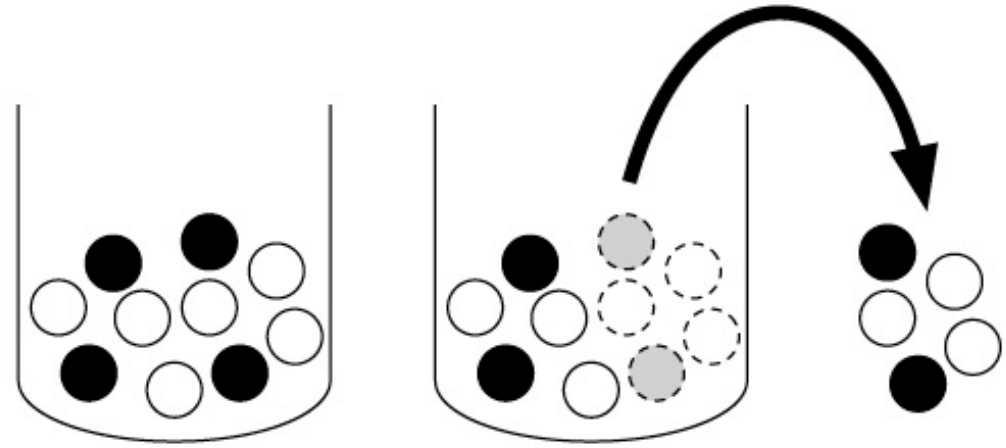

Figure 5: An urn contains $w = 6$ white balls and $b = 4$ black balls. We sample $n = 5$ without replacement. The number X of white balls in the sample is Hypergeometric; here we observe $X = 3$.

Theorem (Hypergeometric PMF)**.** *Consider an urn with w white balls and b blacks balls. We draw n balls out of the urn at random without replacement such that all the* $\binom{w+b}{n}$ *samples are equally likely. Let X be the number of white balls in the sample. Then X is said to have the Hypergeometric distribution with parameters w, b and n:* $X \sim \mathsf{HGeom}(w, b, n)$*. Then the PMF of X is*

$$\mathsf{P}(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for all integers k satisfying $0 \leq k \leq w$ *and* $0 \leq n - k \leq b$*, and* $\mathsf{P}(X = k) = 0$ *otherwise.*

The Hypergeometric distribution comes up in many scenarios in which items in a population are classified using two sets of tags: in the urn story, each ball is either white or black (this is the first set of tags), and each ball is either sampled or not sampled (this is the second set of tags). Furthermore, at least one of these sets of tags is assigned completely at random (in the urn story, the balls are sampled randomly, with all sets of the correct size equally likely). Then $X \sim \mathsf{HGeom}(w, b, n)$ represents the number of twice-tagged items: in the urn story, balls are both white and sampled.

Example: Elk capture-recapture

A forest has N elk. Today, m of the elk are captured, tagged, and released into the wild. At a later date, n elk are recaptured at random. Assume that the recaptured elk are equally likely to be any set of n of the elk, e.g., an elk that has been captured does not learn how to avoid being captured again. The number of tagged elk in the recaptured sample has the $\mathsf{HGeom}(m, N - m, n)$ distribution. The m tagged elk in this story correspond to the white balls and the $N - m$ untagged elk correspond with the black balls. Instead of sampling n balls from the urn, we recapture n elk from the forest.

Theorem. *If $X \sim \mathsf{HGeom}(w, b, n)$ and $Y \sim \mathsf{HGeom}(n, w+b-n, w)$, then X and Y have the same distribution.*

Note (Binomial vs. Hypergeometric). The Binomial and Hypergeometric distributions are often confused. Both are discrete distributions taking on integer values between 0 and n for some n, and both can be interpreted as the number of successes in n Bernoulli trials (for the Hypergeometric, each tagged elk in the recaptured sample can be considered a success and each untagged elk a failure). However, a crucial part of the Binomial story is that the Bernoulli trials involved are *independent*. The Bernoulli trials in the Hypergeometric story are *dependent*, since the sampling is done without replacement: knowing that one elk in our sample is tagged decreases the probability that the second elk will also be tagged.

Theorem. *If $X \sim \mathsf{Bin}(n, p)$, $Y \sim \mathsf{Bin}(m, p)$, and X is independent of Y, then the conditional distribution of X given $X + Y = r$ is $\mathsf{HGeom}(n, m, r)$.*

Theorem (Binomial as a limiting case of the Hypergeometric)**.** *If $X \sim \mathsf{HGeom}(w, b, n)$ and $N = w + b \to \infty$ such that $p = w/(w + b)$ remains fixed, then the PMF of X converges to the $\mathsf{Bin}(n, p)$ PMF.*

4 The discrete uniform distribution

Let C be a finite, nonempty set of numbers. Choose one of these numbers uniformly at random (i.e., all values in C are equally likely). Call the chosen number X. Then X is said to have the *discrete uniform distribution* with parameter C, i.e. $X \sim \mathsf{DUnif}(C)$. The PMF of $X \sim \mathsf{DUnif}(C)$ is

$$\mathsf{P}(X = x) = \frac{1}{|C|},$$

for $x \in C$, and 0 otherwise. Furthermore, for any $A \subseteq C$, we have

$$\mathsf{P}(X \in A) = \frac{|A|}{|C|}.$$

5 Cumulative distribution functions

Definition. The *cumulative distribution function* (CDF) of an random variable X is the function F_X given by $F_X(x) = \mathsf{P}(X \leq x)$.

Theorem (Valid CDFs)**.** *Any CDF F has the following properties.*

- *Increasing: If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.*
- *Right-continuous: As in Figure 5, the CDF is continuous except possibly for having some jumps. Wherever there is a jump, the CDF is continuous from the right. That is, for any a, we have*

$$F(a) = \lim_{x \to a^+} F(x).$$

- *Convergence to 0 and 1 in the limits:*

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$$

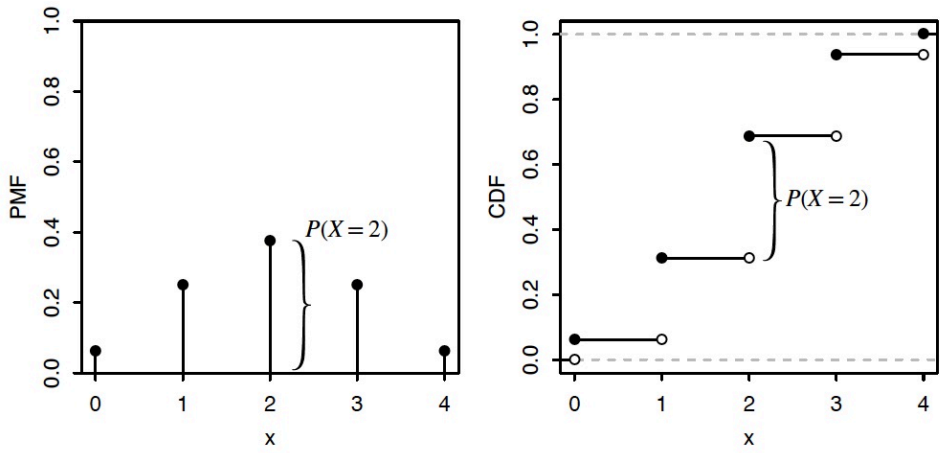

Figure 6: Bin(4, 1/2) PMF (left panel) and CDF (right panel). The height of the vertical bar $\mathsf{P}(X = 2)$ in the PMF is also the height of the jump in the CDF at 2.

6 Functions of random variables

A function of a random variable is also a random variable. That is, if X is a random variable, then X^2, e^X, and $\sin(X)$ are also random variables, as is $g(X)$ for any function $g : \mathbb{R} \to \mathbb{R}$. See Figure 7.

Definition (Function of an random variable)**.** For an experiment with a sample space S, an random variable X, and a function $g : \mathbb{R} \to \mathbb{R}$, $g(X)$ is the random variable that maps any $s \in S$ to $g(X(s))$.

If $g(\cdot)$ is a one-to-one function, finding the PMF of $Y = g(X)$ if we know the PMF of X is straightforward: the support of Y is the set of all $g(x)$ with x in the support of X, and

$$\mathsf{P}(Y = g(x)) = \mathsf{P}(g(X) = g(x)) = \mathsf{P}(X = x).$$

If $g(\cdot)$ is not one-to-one, then for a given y, there may be multiple values of x such that $g(x) = y$. To compute $\mathsf{P}(g(X) = y)$, we need to sum up the probabilities of X taking on any of these candidate values of x.

Theorem (PMF of $g(X)$)**.** *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$. Then the support of $g(X)$ is the set of all y such that $g(x) = y$ for at least one x in the support of X, and the PMF of $g(X)$ is*

$$\mathsf{P}(g(X) = Y) = \sum_{\{x:g(x)=y\}} \mathsf{P}(X = x).$$

for all y in the support of $g(X)$.

Definition (Function of two random variables)**.** Given an experiment with sample space S, if X and Y are random variables that map $s \in S$ to $X(s)$ and $Y(s)$ respectively, then $g(X, Y)$ is the random variable that maps s to $g(X(s), Y(s))$.

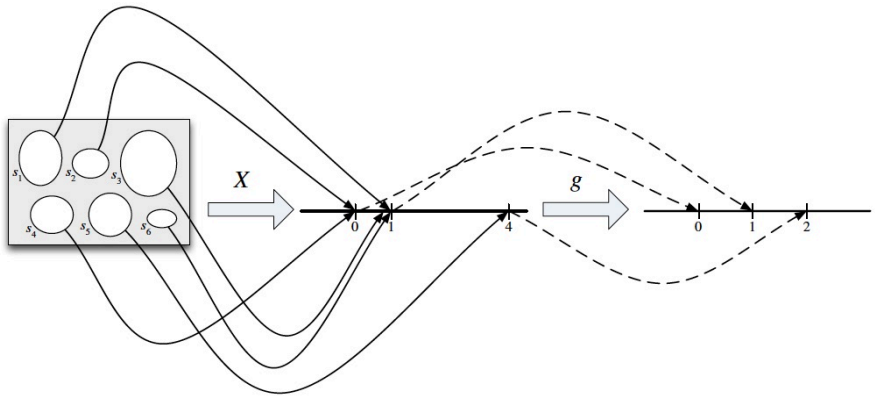

Figure 7: The random variable X is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The function $g(\cdot)$ is the square root function. Composing X and g gives the random variable $g(X) = \sqrt{X}$, which has possible values 0, 1 and 2.

Example: Random walk

A particle moves n steps on a number line. The particle starts at 0, and at each step it moves 1 unit to the right or to the left, with equal probabilities. Assume all steps are independent. Let Y be the particle's position after n steps. We need to find the PMF of Y.

Consider each step to be a Bernoulli trial, where right is considered a success and left is considered a failure. Then the number of steps the particle takes to the right is a $X \sim \mathsf{Bin}(n, 1/2)$ random variable. If $X = j$, then the particle has taken j steps to the right and $n - j$ steps to the left, giving a final position of $j - (n - j) = 2j - n$. So we can express Y as a one-to-one function of X, namely, $Y = 2X - n$. Since X takes values in $\{0, 1, \ldots, n\}$, Y takes values in $\{-n, 2 - n, 4 - n, \ldots, n\}$.

The PMF of Y can be found from the PMF of X:

$$\mathsf{P}(Y = k) = \mathsf{P}(2X - n = k) = \mathsf{P}(X = (n + k)/2) = \binom{n}{\frac{n+k}{2}} \left(\frac{1}{2}\right)^n,$$

if k is an integer between $-n$ and n such that $n + k$ is an even number.

Next, we let $D = |Y|$ be the particle's distance from the origin after n steps. Assume n is even. The function $g(x) = |x|$ is not one-to-one. We have $\{D = 0\} = \{Y = 0\}$, but $\{D = k\} = \{Y = k\} \cup \{Y = -k\}$ for $k = 2, 4, \ldots, n$. So the PMF of D is

$$\begin{aligned} \mathsf{P}(D = 0) &= \binom{n}{\frac{n}{2}} \left(\frac{1}{2}\right)^n, \\ \mathsf{P}(D = k) &= \mathsf{P}(Y = k) + \mathsf{P}(Y = -k) = 2\binom{n}{\frac{n+k}{2}} \left(\frac{1}{2}\right)^n. \end{aligned}$$

7 Independence of random variables

If two random variables X and Y are independent, then knowing the value of X gives no information about the value of Y, and vice versa.

Definition (Independence of two random variables)**.** Random variables X and Y are said to be *independent* if

$$\mathsf{P}(X \leq x, Y \leq y) = \mathsf{P}(X \leq x)\mathsf{P}(Y \leq y),$$

for all $x, y \in \mathbb{R}$. In the discrete case, this is equivalent to the condition

$$\mathsf{P}(X = x, Y = y) = \mathsf{P}(X = x)\mathsf{P}(Y = y),$$

for all x in the support of X and all y in the support of Y.

Definition (Independence of many random variables)**.** Random variables $X_1, \ldots, X_n$ are *independent* if

$$\mathsf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathsf{P}(X_1 \leq x_1) \ldots \mathsf{P}(X_n \leq x_n).$$

for all $x_1, \ldots, x_n \in \mathbb{R}$. For infinitely many random variables, we say that they are independent if every finite subset of the random variables is independent.

If $X_1, \ldots, X_n$ are independent, then they are pairwise independent (i.e., X_i is independent of X_j for $i \neq j$). However, the converse is not true: pairwise independence does not imply independence.

If X and Y are independent, then any function of X alone is independent of any function of Y alone.

Example: Dependent random variables
In a roll of two fair dice, if X is the number on the first die and Y is the number on the second die, then $X + Y$ is not independent of $X - Y$. To see why, remark that

$$\mathsf{P}(X + Y = 12, X - Y = 1) = 0 \neq \mathsf{P}(X + Y = 12)\mathsf{P}(X - Y = 1) = \frac{1}{36}\frac{5}{36}.$$

This means that $X + Y$ and $X - Y$ are dependent.

Definition (Independent and identically distributed (i.i.d.))**.** Variables that are independent and have the same distribution are called random variables independent and identically distributed or, for short, i.i.d.

Random variables are independent if they provide no information about each other. They are identically distributed if they have the same CDF. Whether two random variables are independent has nothing to do with whether or not they have the same distribution.

Theorem. *if $X \sim \mathsf{Bin}(n, p)$, viewed as the number of successes in n independent Bernoulli trials with success probability p, then we can write $X = X_1 + \ldots + X_n$ where the X_i are i.i.d.* $\mathsf{Bern}(p)$.

Theorem. *If $X \sim \mathsf{Bin}(n, p)$, $Y \sim \mathsf{Bin}(m, p)$, and X is independent of Y, then $X + Y \sim \mathsf{Bin}(n + m, p)$.*

Proof.

$$\begin{aligned}
\mathsf{P}(X + Y = k) &= \sum_{j=0}^{k} \mathsf{P}(X + Y = k \mid X = j)\mathsf{P}(X = j), \\
&= \sum_{j=0}^{k} \mathsf{P}(Y = k - j \mid X = j)\mathsf{P}(X = j), \\
&= \sum_{j=0}^{k} \mathsf{P}(Y = k - j)\mathsf{P}(X = j).
\end{aligned}$$

□

Definition (Conditional independence of random variables)**.** Random variables X and Y are *conditionally independent* given another random variable Z if for all $x, y \in \mathbb{R}$ and all z in the support of Z,

$$\mathsf{P}(X \leq x, Y \leq y \mid Z = z) = \mathsf{P}(X \leq x \mid Z = z)\mathsf{P}(Y \leq y \mid Z = z).$$

For discrete random variables, an equivalent definition is to require

$$\mathsf{P}(X = x, Y = y \mid Z = z) = \mathsf{P}(X = x \mid Z = z)\mathsf{P}(Y = y \mid Z = z).$$

Definition (Conditional PMF)**.** For any discrete random variables X and Z, the function $\mathsf{P}(X = x \mid Z = z)$, when considered as a function of x for fixed z, is called the *conditional PMF of X given $Z = z$*.

Independence of random variables does not imply conditional independence, nor vice versa.