1 Introduction to probability and random variables

When an experiment takes place for which different outcomes may occur, but it is not known in advance which one it will be, we can assign *probabilities* of the various outcomes, in the sense that if the same experiment was duplicated over and over, independently, then the probability of a particular outcome can be viewed as the *frequency (proportion) of the experiments for which that outcome occurred*.

For example, when flipping a "fair" coin, we would assign probability 1/2 to it landing heads (H) or tails (T): If we flipped the coin a very large number of times, then the proportion of flips landing H and T would be 1/2 each. In fact if we were to let the number of flips n tend to ∞, then the proportion would converge exactly to 1/2, a consequence of a famous result known as the *Strong Law of Large Numbers* (SLLN).

Similarly, if our experiment is rolling a dice, then each of the 6 sides would be equally likely to occur (probability 1/6). We can even imagine a special dice with k (any integer number) sides that is equally likely to land on any of the k sides when rolled (probability $1/k$).

In many experiments, it is useful to assign a value to each outcome, and that is the essence of a *random variable* (rv) which we usually denote by capital letters such as X, Y, Z and so on.

For example, when flipping a coin, we can define a rv X to take on value 1 if the coin lands H and 0 if it lands T:

$$X = \begin{cases} 1 & \text{if coin lands H,} \\ 0 & \text{if coin lands T.} \end{cases}$$

Then we write $p(1) = P(X = 1) = 1/2$ to denote that "the probability that the random variable X will take on value 1 equals 1/2"; similarly $p(0) = P(X = 0) = 1/2$.

If we were to flip the coin n times, then we would obtain n rvs, $X_1, X_2, \ldots, X_n$, giving us the n outcomes, and each would satisfy $P(X_i = 1) = 1/2 = P(X_i = 0)$, $1 \le i \le n$. The rvs here are also *independent* by which we mean that regardless of what the outcomes are for the first $m < n$ flips (e.g., knowing these outcomes), the rv X_n will still satisfy $P(X_n = 1) = P(X_n = 0) = 1/2$. We thus say that the sequence of rvs $\{X_i : 1 \le i \le n\}$ are *independent and identically distributed* (iid).

1.1 Bernoulli p distribution

The word *distributed* when mentioning independent and identically distributed (iid) in the previous Section refers to the *probability distribution* on the numbers $\{0, 1\}$: $p(1) \stackrel{\text{def}}{=} P(X = 1) = 1/2 = p(0) \stackrel{\text{def}}{=} P(X = 0)$, and $p(0) + p(1) = 1$. This particular distribution is called the *Bernoulli* distribution with "success" probability $p = p(1) = 1/2$. Daniel Bernoulli, from whom the name comes, was a famous Dutch mathematician from the 1700s.

The idea is that we imagine that a coin landing H is a success (given value 1) and it landing tails is a failure (given value 0).

But we can generalize this to allow for any value $0 < p < 1$ for the probability of success, $p = p(1) = P(X = 1)$; and then we have $q = p(0) = P(X = 0) = 1 - p$. Then we would say that X has a Bernoulli(p) distribution, or say that X is a Bernoulli(p) rv.

Let us denote a Bernoulli(p) rv by B for simplicity in what follows; $P(B = 1) = p$, $P(B = 0) = q = 1 - p$. One could imagine this arising from a coin that has been modified to be unfair, but there are many more realistic examples:

1. Let $B = 1$ if you win the lottery today, let $B = 0$ if you do not win today. Here we would expect p to be much smaller than $1/2$; close to 0. For example, if winning the lottery consists of you matching 3 numbers (n_1, n_2, n_3) in exact order, with each number chosen between 0 and 9, then $p = (1/10)^3 = 1/1000$. If it is 5 numbers $(n_1, n_2, n_3, n_4, n_5)$ in exact order, then p decreases to $(1/10)^5 = 1/100,000$

2. An election is about to take place for USA President, and there are two candidates 1 and 2. You go out and ("randomly") choose a voter. Let $B = 1$ if they say they will vote for 1, and let $B = 0$ if they say they will vote for 2. This is the essence of polling, in which we would randomly select a large number n of voters so as to get an estimate of the value of p: We would obtain an iid sequence of such voter outcome rvs $\{B_i : 1 \leq i \leq n\}$, and letting
$$N(n) = \sum_{i=1}^{n} B_i,$$
denote the total number of them that $= 1$, we would estimate
$$p \approx N(n)/n,$$
the proportion of the n voters who said they would vote for candidate 1.

3. $B = 1$ if the price per share of a given stock goes up tomorrow from what it was at closing time today, and $B = 0$ if the price per share of the given stock goes down tomorrow from what it was at closing time today. $p = P(B = 1)$ is thus the probability that the stock will go up.

4. You flip a (fair) coin twice. If we define $B = 1$ if the coin landed H both times, and $B = 0$ if not, then $p = p(1) = (1/2)(1/2) = 1/4$; $p(0) = 1 - p = 3/4$.

5. You roll a pair of dice. Let $B = 1$ if the sum $= 8$, $B = 0$ if not. $p = p(1) = P(B = 1) = 5/36$; $p(0) = 1 - p = 31/36$: There are 36 equally likely possible outcomes for the pair; $\{(i, j) : 1 \leq i \leq 6,\ 1 \leq j \leq 6\}$. Of those, the outcome (event) $\{B = 1\}$ is the same as $\{(4, 4,),\ (3, 5),\ (5, 3),\ (2, 6),\ (6, 2)\}$ which contains 5 of the 36 outcomes.

Whenever we independently perform a sequence of experiments yielding an independent identically distributed sequence of Bernoulli (p) rvs, $B_1, B_2, \ldots$, we say that we have *performed a sequence of iid Bernoulli (p) trials.*

1.2 Binomial (n, p) distribution

If we were to perform a sequence of iid Bernoulli (p) trials (such as flipping a coin n times) in which the probability of success is p, and we let $B_i = 1$ if the i^{th} trial is a success $B_i = 0$ if it is a failure, $1 \leq i \leq n$, then the rv
$$X = \sum_{i=1}^{n} B_i,$$

denotes the "total number of successes out of n trials". In general the outcome of the n values of the B_i is a sequence of 0s and 1s of length n such as
$(0,0,1,1,1,0,1,0,1,0,1,1,1,0,0)$, where here $n=15$. X sums up the number of 1s; $X=8$ in this example.

Note that X here can only take on the values within the set $\{0,1,\ldots,n\}$. X is said to have a *binomial* (n,p) distribution; we will derive its probability distribution next, that is, the values for $p(k)=P(X=k),\ 0\le k\le n$; $p(0)+p(1)+\cdots+p(n)=1$. It is immediate that $p(n)=P(X=n)=p^n$, because this means that all n Bernoulli rvs must be a success. Similarly, $p(0)=P(X=0)=q^n=(1-p)^n$, because this means that all n Bernoulli rvs must be a failure. To obtain the intermediary probabilities, $p(k)=P(X=k),\ 1\le k\le n-1$: If we let $\binom{n}{k}$ denote the number of committees of size k that can be formed from a group of n people, with $1\le k\le n$, it is known (and not too hard to prove) that

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}, \tag{1}$$

where $n!\stackrel{\text{def}}{=}n(n-1)(n-2)\cdots 1$, called a *permutation on* n, is in fact the total number of different ways that you can seat (permute) n people in n chairs. (We also define $0!=1$ in what follows.) The idea here is that by merely re-seating (re-ordering/permuting) the k committee members chosen, we have not changed the committee; it is the same committee, so we don't want to over count.

A little thought reveals that $\binom{n}{k}$ is also the number of different ways (outcomes) that the n Bernoulli trials could yield exactly k successes (and hence exactly $n-k$ failures): If we view the n trials as n people, and the k as the k members to form our committee, then it is immediate. The $k!$ is the number of ways we could re-arrange the ordering of the k successes, while $(n-k)!$ is the number of ways we could re-arrange the ordering of the remaining $n-k$ (the failures). Note that forming a committee of size k also yields a committee of size $n-k$ and visa versa: Each time you form a committee, the remaining people form yet another committee. That is why $\binom{n}{k}=\binom{n}{n-k}$ (recall formula (1).

For example, if $n=4$ and $k=2$, then there are 6 size 2 committees:
$\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}$.
For example committee $\{2,3\}$ means that out of the 4 trials, B_1,B_2,B_3,B_4, trials 2 and 3 were a success, $B_2=B_3=1$, and the other 2 were failures, $B_1=B_4=0$; this yields the outcome sequence $(B_1,B_2,B_3,B_4)=(0,1,1,0)$. We really have two committees when we choose $\{2,3\}$: $\{2,3\}$ and $\{1,4\}$ (the remaining elements). Each of the above 6 outcomes (committees) has the same probability of occurring: $p^2(1-p)^2$; p^2 is the probability of the two successes occurring, and $(1-p)^2$ is the probability of the two failures occurring. Thus, summing up all six yields $P(X=2)=6p^2(1-p)^2$.

In general then when X is binomial (n,p),

$$p(k)=P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, 0\le k\le n, \tag{2}$$

where $p^k(1-p)^{n-k}$ is the probability of any one of the $\binom{n}{k}$ outcomes (committees) containing exactly k successes and $(n-k)$ failures. This probability distribution on the set of numbers $\{0,1,\ldots,n\}$ is called the *binomial* (n,p) *distribution*, $\binom{n}{k}$ are called the *binomial coefficients*, and X is called a *binomial* (n,p) *rv.*

Note that when $p=1/2$, the "fair" coin toss case, then each individual outcome of the n trials has the same probability $p^k(1-p)^{n-k}=(1/2)^k(1/2)^{n-k}=(1/2)^n=1/2^n$, regardless of the value of k, which is an example of "equally likely" probabilities: There are 2^n total possible outcomes (sequences of 0s and 1s) when performing n iid Bernoulli trials, $(B_1, B_2, \ldots, B_n)$, and when $p = 1/2$, they are equally likely. (When $p \neq 1/2$ this is no longer so). In this $p = 1/2$ case $P(X = k) = \binom{n}{k}/2^n$ which says that

> *The probability of exactly k successes out of n trials (when $p = 1/2$) is equal to the number of ways (outcomes for which) there can be exactly k successes, divided by the total number of possible outcomes.*

Examples

1. Consider a coin that lands H with probability $p = 3/4$ and tails with probability $q = 1 - p = 1/4$. You flip it twice, and let X denote the number flips that land H. $n = 2$ and $p = 3/4$. The corresponding binomial $(2, 3/4)$ rv X is thus of the form

$$X = B_1 + B_2,$$

where the B_1, B_2 are independent Bernoulli $(3/4)$ rvs and the $2^2 = 4$ possible outcomes for them are $(0,0),\ (1,0),\ (0,1),\ (0,0)$ yielding the binomial probabilities for $X = B_1 + B_2$ as

$$p(k) = P(X = k) = \begin{cases} 1(1/4)^2 = 1/16, & k = 0, \\ 2(3/4)^1(1/4)^1 = 6/16, & k = 1, \\ 1(3/4)^2 = 9/16, & k = 2. \end{cases}$$

Note that $\binom{2}{0} = 1,\ \binom{2}{1} = 2,\ \binom{2}{2} = 1.$
For a general value of $0 < p < 1$,

$$p(k) = P(X = k) = \begin{cases} (1-p)^2, & k = 0, \\ 2p(1-p), & k = 1, \\ p^2, & k = 2. \end{cases}$$

2. Suppose we flip the coin 3 times. Then the three B_1, B_2, B_3 Bernoulli rvs have the $2^3 = 8$ possible outcomes
$(0,0,0),\ (1,0,0),\ (0,1,0),\ (0,0,1),\ (1,1,0),\ (1,1,0),\ (1,0,1),\ (1,1,1)$, yielding the binomial probabilities for $X = B_1 + B_2 + B_3$ as

$$p(k) = P(X = k) = \begin{cases} 1(1/4)^3 = 4/256, & k = 0, \\ 3(3/4)^1(1/4)^2 = 36/256, & k = 1, \\ 3(3/4)^2(1/4) = 108/256, & k = 2, \\ 1(3/4)^3 = 108/256, & k = 3. \end{cases}$$

Note that $\binom{3}{0} = 1,\ \binom{3}{1} = \binom{3}{2} = 3,\ \binom{3}{3} = 1.$
For a general value of $0 < p < 1$,

$$p(k) = P(X = k) = \begin{cases} (1-p)^3, & k = 0, \\ 3p(1-p)^2, & k = 1, \\ 3p^2(1-p), & k = 2, \\ p^3, & k = 3. \end{cases}$$

In general, computing binomial probabilities by hand when n gets large becomes impractical; a computer program can be used (they are even built into most computer software programs) and tables are even available in any basic probability and statistics book. There are even very nice approximations to the probabilities when n is sufficiently large (using the normal distribution; this is a special case of the *Central Limit Theorem* due to the French mathematician Abraham de Moivre in the 1730s), or when n is large and p is small (then one can use the *Poisson* distribution, name after the French mathematician Siméon Denis Poisson in the 1830s).

1.3 The expected value (mean) of a random variable

If B is Bernoulli(p), then its expected value, denoted by $E(B)$ is simply the average of the values it takes on

$$E(B) = 1p(1) + 0p(0) = p.$$

More generally, if X is a random variable with distribution (say) $P(X = k) = p(k),\ 0 \leq k \leq n$, then

$$E(X) = 0p(0) + 1p(1) + \cdots + np(n).$$

The way to think about this: If we took a large number N independent copies of X, $X_1, \ldots, X_N$ and average them, then we would obtain exactly $E(X)$ as $N \to \infty$: Letting

$$\overline{X}(N) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} X_i,$$

we have that (with probability equal to one)

$$\lim_{N\to\infty} \overline{X}(N) = E(X).$$

An easy nice property of expected values is linearity:
If c is a constant and X and Y are rvs, then $E(cX + Y) = cE(X) + E(Y)$.
Thus, for example, when

$$X = \sum_{i=1}^{n} B_i,$$

is a binomial (n, p) rv, then

$$E(X) = \sum_{i=1}^{n} E(B_i) = \sum_{i=1}^{n} p = np.$$

This makes good sense: each trial is successful with probability p so the average number of successes out of n such trials is np.
Another nice property concerns the expected value of independent products:
If X and Y are independent rvs, then $E(XY) = E(X)E(Y)$.

1.4 The continuous uniform distribution on the interval $(0, 1)$.

If we were to place n (a large integer) equally spaced points within the continuous interval $[0, 1]$, $1/n, 2/n, \ldots, n/n = 1$, and imagine a dice with n sides (faces) that are equally likely to be the side that the dice lands on when rolled, then we can define a random variable U_n via $U_n = i/n$ if the dice lands on side i; it takes on any of the n values $i/n,\ 1 \leq i \leq n$, with probability $1/n$. As n gets larger and larger, the probability of U_n taking on any specific such value tends to 0 because i/n tends to 0 as $n \to \infty$. It thus makes more sense to consider intervals into which U_n takes values.

Note that for $1 \leq i < j \leq 1$, the probability that $i/n \leq U_n \leq j/n$, written as $P(i/n \leq U_n \leq j/n)$, or equivalently as $P(U_n \in [i/n, j/n])$, is just the sum $(j-i)/n$, since there are $j-i$ values, $i/n, (i+1)/n, \ldots, j/n$, that are contained in the interval $[i/n, j/n]$ and each has probability $1/n$. Note further that $(j-i)/n$ is simply the length of the interval $[i/n, j/n]$, so *any* interval I of length $(j-i)/n$ of this kind has the property that $P(U_n \in I) = (j-i)/n$. As $n \to \infty$, so that the mesh of points $1/n, 2/n, \ldots, n/n$ approaches a continuum, we intuitively see that we would obtain a limiting random variable U with the property that $P(a \leq U \leq b) = b - a$ for any $0 < a < b < 1$. We also see that $P(U = x) = 0$ for any particular x because $1/n \to 0$ as $n \to \infty$. Thus we also have $P(a < U < b) = P(a \leq U < b) = P(a < U \leq b) = a - b$ as well; the end points have no effect since they have probability 0 of being selected.

The probability distribution on $(0, 1)$ that we obtain is called the continuous *uniform distribution over the interval* $(0, 1)$. We typically denote a rv with this distribution by U and call it a uniformly distributed rv over $(0, 1)$. It is uniquely defined by having the property that $F(x) \stackrel{\text{def}}{=} P(U \leq x) = x$ for any $0 \leq x \leq 1$. (Just set $a = 0,\ b = x$.) Also, $P(U > x) = 1 - P(U \leq x) = 1 - x$.

As a function of x, this function $F(x)$ is called the *cumulative distribution function* of U.

U represents a continuous version of the *discrete uniform distribution* over a set of n points, in which each point has equal probability of being chosen, $1/n$. For example, if n playing cards, labeled $1 - n$, are shuffled, and you choose one, you are equally likely to have chosen any one of the n cards.

1.5 Simulating random variables from your computer

It turns out that a computer can hand you, upon demand, copies of continuous uniform rvs U, denoted by $U_1, U_2, \ldots$ and such that they are independent. It uses its random number generator to do so; we will discuss the mechanism (algorithms) of how it does this some time later on. (The computer generated numbers are not really truly random, they are what is called *pseudo-random*; but for our purposes we will not worry about that for now.)

Being able to 'generate' from a computer independent copies of U is the starting point of *stochastic simulation*. As we shall see, if we can generate uniforms U, then we can generate ('simulate') rvs X with any desired probability distribution. To see what we have in mind consider the following examples:

1. *Simulating a copy B of a Bernoulli(p) rv.* If we want such a B we can do so with the following simple algorithm:

   (i) Enter p.

   (ii) Generate a U

   (iii) Set

$$B = \begin{cases} 1 & \text{if } U \leq p\ , \\ 0 & \text{if } U > p. \end{cases}$$

This works because $P(B = 1) = P(U \leq p) = F(p) = p$, and $P(B = 0) = P(U > p) = 1 - F(p) = 1 - p$ exactly as is required. Indeed B has the Bernoulli (p) distribution as desired. Repeating this algorithm's steps (ii)-(iii) using independent copies of U then yields independent copies of Bernoulli (p) rvs, say $B_1, B_2, \ldots$

In Python, here is the code:

Python Algorithm for generating a Bernoulli (p) **random variable** B

```
import random
#Bernoulli (p) algorithm
def Bernoulli(p):
    #Generate a uniform RV U
    U = random.random()
    #Check if it is <=p
    if U <=p:
        return 1
    else:
        return 0
```

2. *Simulating a copy X of a Binomial(n, p) rv.* Simply generate n independent copies of B, denoted by $B_1, \ldots, B_n$ using the algorithm above and set

$$X = \sum_{i=1}^{n} B_i.$$

Python Algorithm for generating a Binomial(n, p) **rv random variable** X

```
def binomial(n,p):
    total = 0
    #Perform n independent Bernoulli (p) trials and sum them up
    for k in range(n):
        total += Bernoulli(p)
    return total
```