EE 178/278A Probabilistic Systems Analysis

Spring 2014 Tse/Hussami Lecture 6

The Coupon Collector's Problem: An Application

We consider a simplified description of the BitTorrent peer-to-peer file-sharing protocol. When someone uploads a file to BitTorrent, it breaks up the file into many chunks, randomly selects many peers in the network, and sends each peer one chunk of the file. When another BitTorrent client wants to download the file, it repeatedly queries a random peer to return its chunk, until the client has all of the chunks of the file. Thus, in each iteration the client obtains another random chunk of the file.

Lets consider an example where a movie is broken down into 10 chunks. Each of you will pretend to be a BitTorrent server and please randomly select a number from 1 to 10. This will be the random chunk that you are storing. I am now querying the servers one by one to download the chunks, so please now shout out your number:

$$6,7,4,1,2,8,9,6,2,3,3,8,9,10,4,4,1,7,4,7,5,8,8$$

The above sequence represents an outcome of our experiment. We are interested in how many servers I need to query before I can download the movie. In this outcome, it is 21. Note that it took many queries before we find the last chunk, 5. If the file consists of n chunks, and in each iteration the client obtains a random chunk, then we can expect that downloaders will have to wait a long time to obtain the whole file: they'll have to perform about $n\log(10n)$ iterations before they have a decent chance of collecting all the chunks. This is inefficient, so in practice BitTorrent is forced to go out of its way to include special mechanisms to speed up downloading: BitTorrent clients download chunks randomly until they have a large fraction of the file, then they switch to searching for the specific chunks they are missing.

The number of servers that have to be queried until collecting all the movie chunks is an example of a *random variable*. This random variable assigns a number to each outcome of the experiment, which in this case is a sequence of servers queried. The general concept of random variable will be our next topic.

Random Variables: Distribution and Expectation[1]

Random Variables

Question: The homeworks of 20 students are collected in, randomly shuffled and returned to the students. How many students receive their own homework?

To answer this question, we first need to specify the probability space: plainly, it should consist of all 20! permutations of the homeworks, each with probability $\frac{1}{20!}$. [Note that this is the same as the probability space for card shuffling, except that the number of items being shuffled is now 20 rather than 52.] It helps to have a picture of a permutation. Think of 20 books lined up on a shelf, labeled from left to right with $1,2,\ldots,20$. A permutation π is just a reordering of the books, which we can describe just by listing their

[1] This section is modified from Berkeley EECS 70 notes.

labels from left to right. Let's denote by π_i the label of the book that is in position i. We are interested in the number of books that are still in their original position, i.e., in the number of i's such that $\pi_i = i$. These are often known as *fixed points* of the permutation.

Of course, our question does not have a simple numerical answer (such as 6), because the number depends on the particular permutation we choose (i.e., on the outcome). Let's call the number of fixed points X. To make life simpler, let's also shrink the class size down to 3 for a while. The following table gives a complete listing of the sample space (of size $3! = 6$), together with the corresponding value of X for each outcome. [We use our bookshelf convention for writing a permutation: thus, for example, the permutation 312 means that book 3 is on the left, book 1 in the center, and book 2 on the right. You should check you agree with this table.]

permutation π	value of X
123	3
132	1
213	1
231	0
312	0
321	1

Thus we see that X takes on values 0, 1 or 3, depending on the outcome. A quantity like this, which takes on some numerical value at each outcome, is called a *random variable* (or *r.v.*) on the sample space.

Definition 6.1 (random variable): A *random variable* X on a sample space Ω is a function that assigns to each outcome $\omega \in \Omega$ a real number $X(\omega)$.

The r.v. X in our permutation example above is completely specified by its values at all outcomes, as given in the above table. (Thus, for example, $X(123) = 3$ etc.)

A random variable can be visualized in general by the picture in Figure 1[2]. Note that the term "random variable" is really something of a misnomer: it is a function so there is nothing random about it and it is definitely not a variable! What is random is which outcome of the experiment is realized and hence the value that the random variable maps the outcome to.

Distribution

When we introduce the basic probability space in Note 1, we defined two things: 1) the sample space Ω consisting of all the possible outcomes of the experiment; 2) the probability of each of the outcomes. Analogously, there are two things important about any random variable: 1) the set of values that it can take ; 2) the probabilities with which it takes on the values. Since a random variable is defined on a probability space, we can calculate these probabilities given the probabilities of the outcomes. Let a be any number in the range of a random variable X. Then the set

$$\{\omega \in \Omega : X(\omega) = a\}$$

is an *event* in the sample space (do you see why?). We usually abbreviate this event to simply "$X = a$". Since $X = a$ is an event, we can talk about its probability, $\mathbf{P}(X = a)$. The collection of these probabilities, for all possible values of a, is known as the *distribution* of the r.v. X.

[2] This and other figures in this note are inspired by figures in Chapter 2 of "Introduction to Probability" by D. Bertsekas and J. Tsitsiklis.

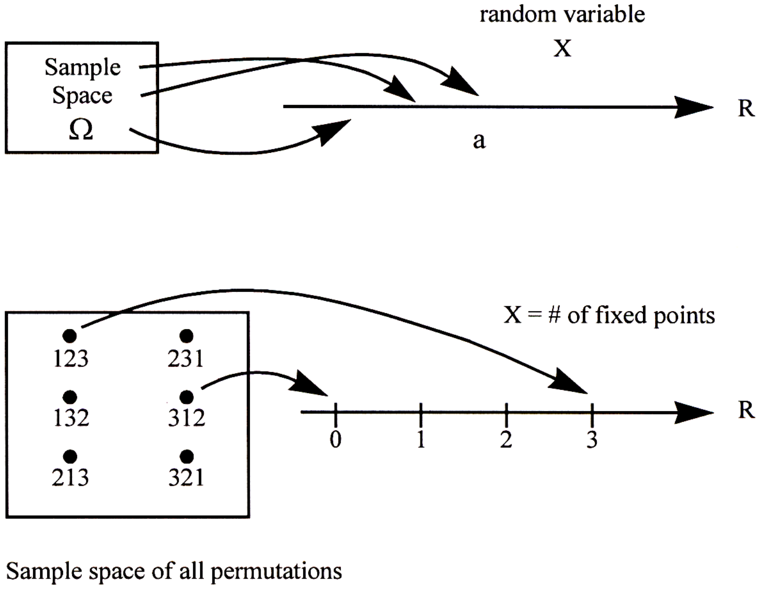

Figure 1: Visualization of how a random variable is defined on the sample space.

Definition 6.2 (distribution or probability mass function): The distribution or the probability mass function of a random variable X is the collection of values $\{(a, \mathbf{P}_X(a) = \mathbf{P}(X = a)) : a \in \mathscr{A}\}$, where $\mathscr{A}$ is the set of all possible values taken by X.

Thus the distribution of the random variable X in our permutation example above is

$$\mathbf{P}(X = 0) = \frac{1}{3}; \qquad \mathbf{P}(X = 1) = \frac{1}{2}; \qquad \mathbf{P}(X = 3) = \frac{1}{6};$$

and $\mathbf{P}(X = a) = 0$ for all other values of a.

The distribution of a random variable can be visualized as a bar diagram, shown in Figure 2. The x-axis represents the values that the random variable can take on. The height of the bar at a value a is the probability $\mathbf{P}(X = a)$. Each of these probabilities can be computed by looking at the probability of the corresponding event in the sample space.

Note that the collection of events $X = a$, $a \in \mathscr{A}$, satisfy two important properties:

- any two events $X = a_1$ and $X = a_2$ with $a_1 \neq a_2$ are disjoint.
- the union of all these events is equal to the entire sample space Ω.

The collection of events thus form a *partition* of the sample space (see Figure 2). Both properties follow directly from the fact that X is a function defined on Ω, i.e. X assigns a unique value to each and every possible outcome in Ω. As a consequence, the sum of the probabilities $\mathbf{P}(X = a)$ over all possible values of a is exactly 1. So when we sum up the probabilities of the events $X = a$, we are really summing up the probabilities of all the outcomes.

Example: The binomial distribution: This is one of the most important distributions in probability. It can be defined in terms of a coin-tossing experiment. Consider n independent tosses of a biased coin with Heads probability p. Each outcome is a sequence of tosses. For example, when $n = 3$, the sample space Ω is $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

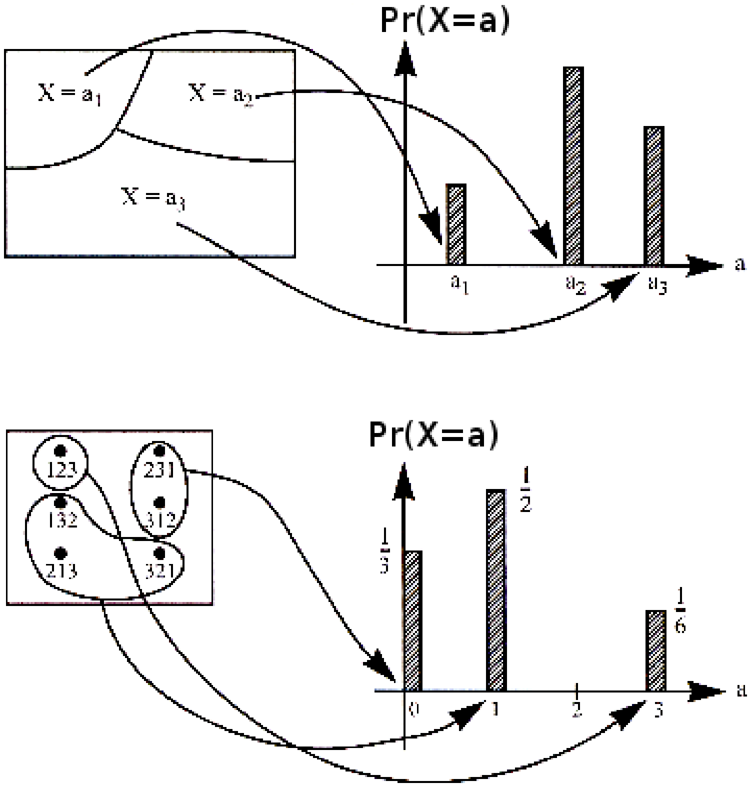

Figure 2: Visualization of how the distribution of a random variable is defined.

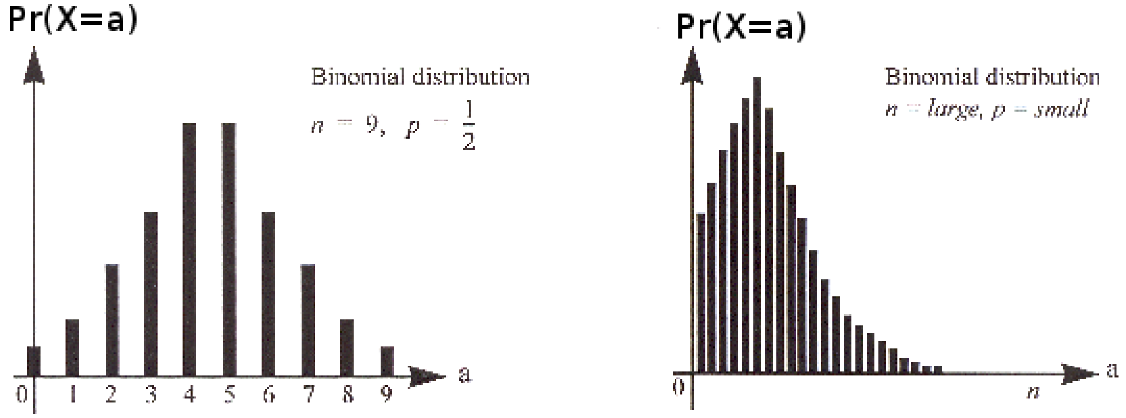

Figure 3: The binomial distributions for two choices of (n, p).

Let X be the number of Heads. Note that this is a function on the sample space: for each outcome ω, $X(\omega)$ is the number of Heads in ω. For example, $X(THH) = 2$ To compute the distribution of X, we first enumerate the possible values X can take on. They are simply $0, 1, \ldots, n$. Then we compute the probability of each event $X = i$ for $i = 0, \ldots, n$. The probability of the event $X = i$ is the sum of the probabilities of all the outcomes with i Heads. Any such outcome has a probability $p^i(1-p)^{n-i}$. There are exactly $\binom{n}{i}$ of these outcomes. So

$$\mathbf{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \qquad i = 0, 1, \ldots n \tag{1}$$

This is the *binomial* distribution with parameters n and p. A random variable with this distribution is called a *binomial* random variable (for brevity, we will say $X \sim \text{Bin}(n, p)$). An example of a binomial distribution is shown in Figure 3.

Although we define the binomial distribution in terms of an experiment involving tossing coins, this distribution is useful for modeling many real-world problems. Consider for example the problem of reliable data storage in the face of hard disk failure. The technology is called RAID. (See http://en.wikipedia.org/wiki/RAID.) Reliability is provided by adding redundancy and using error-correction coding: the data is distributed across n disks and can be recovered as long as no more than k disks fail. (The parameters n and k depend on the level of RAID used.) Assuming each disk fails independently with probability p, the number of disk failures X is binomial distributed with parameters n and p. So the probability of unrecoverability of the data is given by :

$$\mathbf{P}(X > k) = \sum_{i=k+1}^{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

For a given value of p, we can choose k large enough such that this probability is no less than, say, 0.99.