CS 70 Discrete Mathematics and Probability Theory
Summer 2019 Course Notes **Note 15**

Random Variables: Distribution

Recall our setup of a probabilistic experiment as a procedure of drawing a sample from a set of possible values, and assigning a probability for each possible outcome of the experiment. For example, if we toss a fair coin n times, then there are 2^n possible outcomes, each of which is equally likely and has probability $\frac{1}{2^n}$.

Now suppose we want to make a measurement in our experiment. For example, we can ask what is the number of heads in n coin tosses; call this number X. Of course, X is not a fixed number, but it depends on the actual sequence of coin flips that we obtain. For example, if $n = 4$ and we observe the outcome $\omega = HTHH$, then is $X = 3$; whereas if we observe the outcome $\omega = HTHT$, then the is $X = 2$. In this example of n coin tosses, we only know that X is an integer between 0 and n, but we do not know what its exact value is until we observe which outcome of n coin flips is realized and count how many heads there are. Because every possible outcome is assigned a probability, the value X also carries with it a probability for each possible value it can take. The table below lists all the possible values X can take in the example of $n = 4$ coin tosses, along with their respective probabilities.

outcomes ω	value of X (# heads)	probability of occurring
$TTTT$	0	1/16
$HTTT$, $THTT$, $TTHT$, $TTTH$	1	4/16
$HHTT$, $HTHT$, $HTTH$, $THHT$, $THTH$, $TTHH$	2	6/16
$HHHT$, $HHTH$, $HTHH$, $THHH$	3	4/16
$HHHH$	4	1/16

Such a value X that depends on the outcome of the probabilistic experiment is called a *random variable* (abbreviated *r.v.*). As we see from the example above, a random variable X typically does not have a definitive value, but instead only has a probability *distribution* over the set of possible values X can take, which is why it is called random. So the question "What is the number of heads in n coin tosses?" does not exactly make sense because the answer X is a random variable.

1 Random Variables

Before we formalize the above notions, let us consider another example to enforce our conceptual understanding of a random variable.

Example: Fixed Points of Permutations

Question: Suppose we collect the homeworks of n students, randomly shuffle them, and return them to the students. How many students receive their own homework?

Here the probability space consists of all $n!$ permutations of the homeworks, each with equal probability $\frac{1}{n!}$. If we label the homeworks as $1, 2, \ldots, n$, then each sample point is a permutation $\pi = (\pi_1, \ldots, \pi_n)$ where π_i

is the homework that is returned to the i-th student. Note that $\pi_1, \ldots, \pi_n \in \{1, 2, \ldots, n\}$ are all distinct, so each element in $\{1, \ldots, n\}$ appears exactly once in the permutation π.

In this setting, the i-th student receives her own homework if and only if $\pi_i = i$. Then the question "How many students receive their own homework?" translates into the question of how many indices (i's) satisfy $\pi_i = i$. These are known as fixed points of the permutation. As in the coin flipping case above, our question does not have a simple numerical answer (such as 4), because the number depends on the particular permutation we choose (i.e., on the sample point). Let us call the number of fixed points X_n, which is a random variable.

To illustrate the idea concretely, let us consider the example $n = 3$. The following table gives a complete listing of the sample space (of size $3! = 6$), together with the corresponding value of X_3 for each sample point. Here we see that X_3 takes on values 0, 1 or 3, depending on the sample point. You should check that you agree with this table.

permutation π	value of X_3 (# fixed points)
123	3
132	1
213	1
231	0
312	0
321	1

Formal Definition of a Random Variable

We now formalize the concepts discussed above.

Definition 15.1 (Random Variable). *A random variable X on a sample space Ω is a function $X : \Omega \to \mathbb{R}$ that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.*

Until further notice, we will restrict our attention to random variables that are discrete, i.e., they take values in a range that is finite or countably infinite. This means even though we define X to map Ω to $\mathbb{R}$, the actual set of values $\{X(\omega) : \omega \in \Omega\}$ that X takes is a discrete subset of $\mathbb{R}$.

A random variable can be visualized in general by the picture in Figure 1.[1] 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 sample point of the experiment is realized and hence the value that the random variable maps the sample point to.

2 Probability Distribution

When we introduced the basic probability space in Note 13, we defined two things:

1. The sample space Ω consisting of all the possible outcomes (sample points) of the experiment.
2. The probability of each of the sample points.

[1] The figures in this note are inspired by figures in Chapter 2 of *Introduction to Probability* by D. Bertsekas and J. Tsitsiklis.

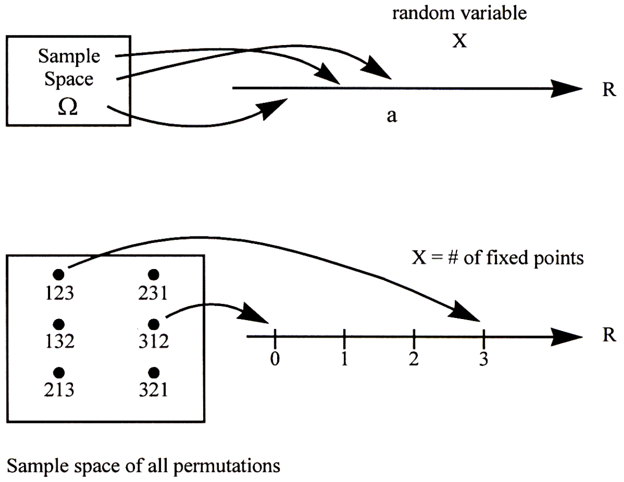

Figure 1: Visualization of how a random variable is defined on the sample space.

Analogously, there are two important things 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 sample points. 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 (simply because it is a subset of Ω). We usually abbreviate this event to simply "$X = a$". Since $X = a$ is an event, we can talk about its probability, $\mathbb{P}[X = a]$. The collection of these probabilities, for all possible values of a, is known as the *distribution* of the random variable X.

Definition 15.2 (Distribution)**.** *The distribution of a discrete random variable X is the collection of values $\{(a, \mathbb{P}[X = a]) : a \in \mathscr{A}\}$, 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:

$$\mathbb{P}[X = 0] = \frac{1}{3}, \qquad \mathbb{P}[X = 1] = \frac{1}{2}, \qquad \mathbb{P}[X = 3] = \frac{1}{6},$$

and $\mathbb{P}[X = a] = 0$ 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 assume. The height of the bar at a value a is the probability $\mathbb{P}[X = a]$. 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 $X = a$, for $a \in \mathscr{A}$, 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.

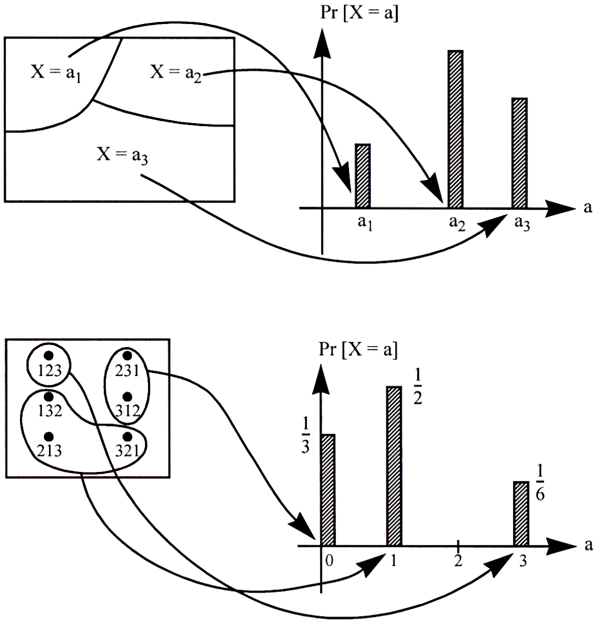

Figure 2: Visualization of how the distribution of a random variable is defined.

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 sample point in Ω. As a consequence, the sum of the probabilities $\mathbb{P}[X = a]$ over all possible values of a is exactly equal to 1. So when we sum up the probabilities of the events $X = a$, we are really summing up the probabilities of all the sample points.

2.1 Bernoulli Distribution

A simple yet very useful probability distribution is the *Bernoulli* distribution of a random variable which takes value in $\{0,1\}$:

$$\mathbb{P}[X = i] = \begin{cases} p, & \text{if } i = 1, \\ 1-p, & \text{if } i = 0, \end{cases}$$

where $0 \leq p \leq 1$. We say that X is distributed as a *Bernoulli* random variable with parameter p, and write

$$X \sim \text{Bernoulli}(p).$$

2.2 Binomial Distribution

Let us return to our coin tossing example above, where we defined our random variable X to be the number of heads. More formally, consider the random experiment consisting of n independent tosses of a biased coin that shows H with probability p. Each sample point ω is a sequence of tosses, and $X(\omega)$ is defined to be the number of heads in ω. For example, when $n = 3$, $X(THH) = 2$.

To compute the distribution of X, we first enumerate the possible values that X can take. They are simply $0,1,\ldots,n$. Then we compute the probability of each event $X = i$ for $i = 0,1,\ldots,n$. The probability of the event $X = i$ is the sum of the probabilities of all the sample points with exactly i heads (for example, if $n = 3$ and $i = 2$, there would be three such sample points $\{HHT,HTH,THH\}$). Any such sample point has probability $p^i(1-p)^{n-i}$, since the coin flips are independent. There are exactly $\binom{n}{i}$ of these sample points.

Hence,

$$\mathbb{P}[X=i]=\binom{n}{i}p^i(1-p)^{n-i}, \qquad \text{for } i=0,1,\ldots,n. \tag{1}$$

This distribution, called the *binomial* distribution, is one of the most important distributions in probability. A random variable with this distribution is called a *binomial* random variable, and we write

$$X \sim \text{Bin}(n,p),$$

where n denotes the number of trials and p the probability of success (observing an H in the example). An example of a binomial distribution is shown in Figure 3. Notice that due to the properties of X mentioned above, it must be the case that $\sum_{i=0}^{n}\mathbb{P}[X=i]=1$, which implies that $\sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i}=1$. This provides a probabilistic proof of the Binomial Theorem!

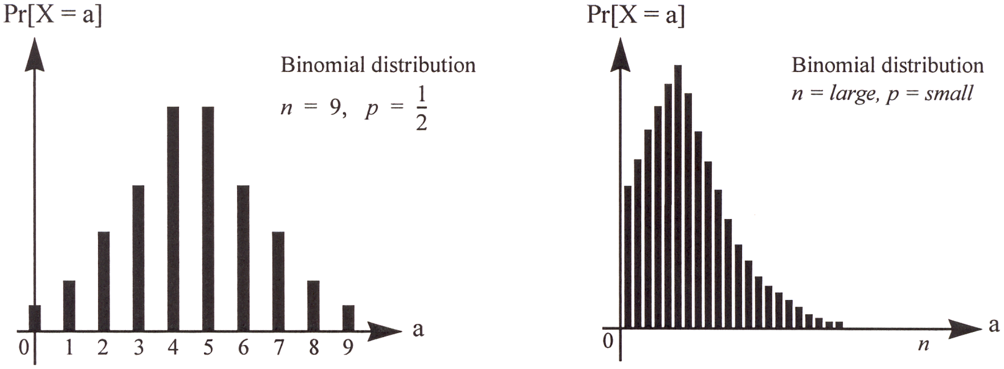

Figure 3: The binomial distributions for two choices of (n,p).

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 error correction problem studied in Note 9. Recall that we wanted to encode n packets into $n+k$ packets such that the recipient can reconstruct the original n packets from any n packets received. But in practice, the number of packet losses is random, so how do we choose k, the amount of redundancy? If we model each packet getting lost with probability p and the losses are independent, then if we transmit $n+k$ packets, the number of packets received is a random variable X with binomial distribution: $X \sim \text{Bin}(n+k, 1-p)$ (we are tossing a coin $n+k$ times, and each coin turns out to be a head (packet received) with probability $1-p$). So the probability of successfully decoding the original data is:

$$\mathbb{P}[X \geq n] \;=\; \sum_{i=n}^{n+k}\mathbb{P}[X=i] \;=\; \sum_{i=n}^{n+k}\binom{n+k}{i}(1-p)^i p^{n+k-i}.$$

Given fixed n and p, we can choose k such that this probability is no less than, say, 0.99.

2.3 Geometric Distribution

Consider repeatedly tossing a biased coin with Heads probability p. Let X denote the number of tosses until the first Head appears. Then X is a random variable that takes values in $\mathbb{Z}^+$, the set of positive integers. The event that $X=i$ is equal to the event of observing Tails for the first $i-1$ tosses and getting Heads in the i-th toss, which occurs with probability $(1-p)^{i-1}p$. Such a random variable is called a geometric random variable.

The geometric distribution frequently occurs in applications because we are often interested in how long we have to wait before a certain event happens: how many runs before the system fails, how many shots before one is on target, how many poll samples before we find a Democrat, how many retransmissions of a packet before successfully reaching the destination, etc.

Definition 15.3 (Geometric Distribution)**.** *A random variable X for which*

$$\mathbb{P}[X=i]=(1-p)^{i-1}p, \qquad \text{for } i=1,2,3,\ldots,$$

is said to have the geometric distribution with parameter p. This is abbreviated as $X\sim$ Geometric(p).

As a sanity check, we can verify that the total probability of X is equal to 1:

$$\sum_{i=1}^{\infty}\mathbb{P}[X=i]=\sum_{i=1}^{\infty}(1-p)^{i-1}p=p\sum_{i=1}^{\infty}(1-p)^{i-1}=p\times\frac{1}{1-(1-p)}=1,$$

where in the second-to-last step we have used the formula for geometric series.

If we plot the distribution of X (i.e., the values $\mathbb{P}[X=i]$ against i) we get a curve that decreases monotonically by a factor of $1-p$ at each step, as illustrated in Figure 4.

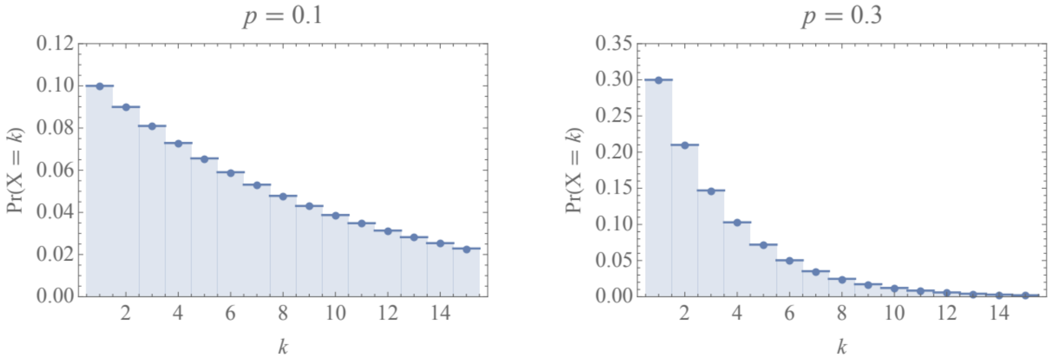

Figure 4: Illustration of the Geometric(p) distribution for $p=0.1$ and $p=0.3$.

2.4 Poisson Distribution

Consider the number of clicks of a Geiger counter, which measures radioactive emissions. The average number of such clicks per unit time, λ, is a measure of radioactivity, but the actual number of clicks fluctuates according to a certain distribution called the Poisson distribution. What is remarkable is that the average value, λ, completely determines the probability distribution on the number of clicks X.

Definition 15.4 (Poisson distribution)**.** *A random variable X for which*

$$\mathbb{P}[X=i]=\frac{\lambda^i}{i!}\mathrm{e}^{-\lambda}, \qquad \text{for } i=0,1,2,\ldots \tag{2}$$

is said to have the Poisson distribution with parameter λ. This is abbreviated as $X\sim$ Poisson(λ).

To make sure this is a valid definition, let us check that (2) is in fact a distribution, i.e., that the probabilities sum to 1. We have

$$\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\mathrm{e}^{-\lambda}=\mathrm{e}^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}=\mathrm{e}^{-\lambda}\times\mathrm{e}^{\lambda}=1.$$

In the second-to-last step, we used the Taylor series expansion $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.

The Poisson distribution is also a very widely accepted model for so-called "rare events," such as misconnected phone calls, radioactive emissions, crossovers in chromosomes, the number of cases of disease, the number of births per hour, etc. This model is appropriate whenever the occurrences can be assumed to happen randomly with some constant density in a continuous region (of time or space), such that occurrences in disjoint subregions are independent. One can then show that the number of occurrences in a region of unit size should obey the Poisson distribution with parameter λ.

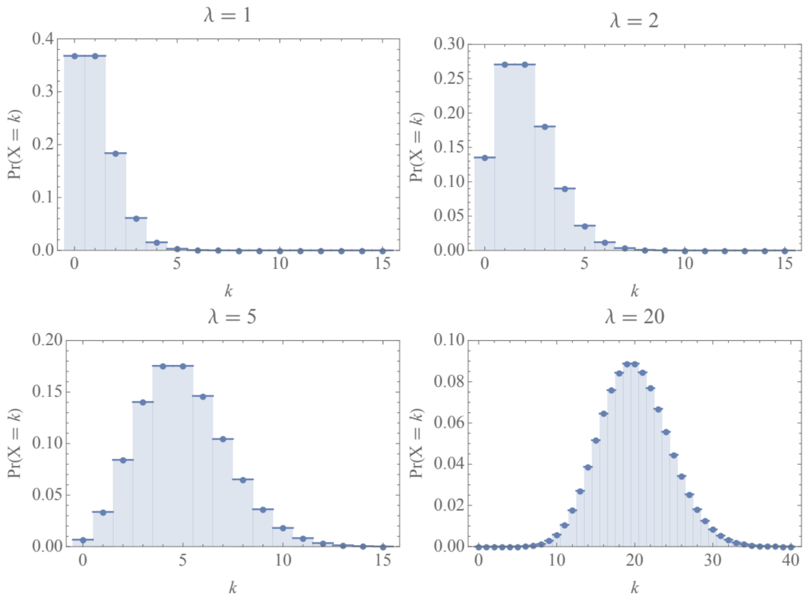

Figure 5: Illustration of the Poisson(λ) distribution for $\lambda = 1, 2, 5, 20$.

The Poisson distribution can also be viewed as the limit of a Binomial distribution. For Summer 2019, this is out of scope of the class. Those who are interested can see Note 19 for more details.

Example

Suppose when we write an article, we make an average of 1 typo per page. We can model this with a Poisson random variable X with $\lambda = 1$. So the probability that a page has 5 typos is

$$\mathbb{P}[X = 5] = \frac{1^5}{5!} e^{-1} = \frac{1}{120\,e} \approx \frac{1}{326}.$$

Now suppose the article has 200 pages. If we assume the number of typos in each page is independent, then the probability that there is at least one page with exactly 5 typos is

$$\begin{aligned}
\mathbb{P}[\exists \text{ a page with exactly 5 typos}] &= 1 - \mathbb{P}[\text{every page has} \neq 5 \text{ typos}] \\
&= 1 - \prod_{k=1}^{200} \mathbb{P}[\text{page } k \text{ has} \neq 5 \text{ typos}] \\
&= 1 - \prod_{k=1}^{200} (1 - \mathbb{P}[\text{page } k \text{ has exactly 5 typos}]) \\
&= 1 - \left(1 - \frac{1}{120\,e}\right)^{200},
\end{aligned}$$

where in the last step we have used our earlier calculation for $\mathbb{P}[X=5]$. □

3 Multiple Random Variables and Independence

Often one is interested in multiple random variables on the same sample space. Consider, for example, the sample space of flipping two coins. One could define many random variables: for example a random variable X indicating the number of heads in a sequence of coin tosses, or a random variable Y indicating the number of tails, or a random variable Z indicating whether the first is H or not. Note that for each sample point, any random variable has a specific value: e.g., for $\omega = HTT$, we have $X(\omega)=1$, $Y(\omega)=2$, and $Z(\omega)=1$.

The concept of a distribution can then be extended to probabilities for the combination of values for multiple random variables.

Definition 15.5. *The joint distribution for two discrete random variables X and Y is the collection of values $\{((a,b),\mathbb{P}[X=a,Y=b]) : a\in\mathscr{A}, b\in\mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X=a]$ for X is called the *marginal distribution* for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X=a] = \sum_{b\in\mathscr{B}} \mathbb{P}[X=a,Y=b].$$

The marginal distribution for Y is analogous, as is the notion of a joint distribution for any number of random variables.

A joint distribution over random variables $X_1,\dots,X_n$ (for example, X_i could be the value of the ith roll of a sequence of n die rolls) is $\mathbb{P}[X_1=a_1,\dots,X_n=a_n]$, where $a_i\in\mathscr{A}_i$ and $\mathscr{A}_i$ is the set of possible values for X_i. The marginal distribution for X_i is simply the distribution for X_i and can be obtained by summing over all the possible values of the other variables, but in some cases can be derived more simply. We proceed to one such case.

Independence for random variables is defined in an analogous fashion to independence for events:

Definition 15.6 (Independence)**.** *Random variables X and Y on the same probability space are said to be* independent *if the events $X=a$ and $Y=b$ are independent for all values a,b. Equivalently, the joint distribution of independent r.v.'s decomposes as*

$$\mathbb{P}[X=a,Y=b] = \mathbb{P}[X=a]\mathbb{P}[Y=b], \quad \forall a,b.$$

Mutual independence of more than two r.v.'s is defined similarly.

A very important example of independent random variables are indicator random variables for independent events. If I_i denotes the indicator r.v. for the i-th toss of a coin being H, then $I_1,\dots,I_n$ are mutually independent random variables. This example motivates the commonly used phrase "*independent and identically distributed (i.i.d.)* set of random variables." In this example, $\{I_1,\dots,I_n\}$ is a set of i.i.d. indicator random variables.